ACS OMEGA

Open Access

http://pubs.acs.org/journal/acsodf

Article

Modeling High Energy Molecules and Screening to Find Novel High Energy Candidates

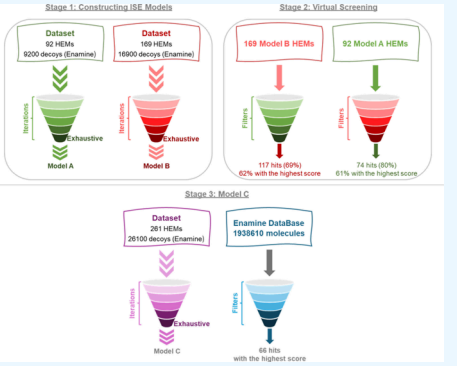
Mazal Rachamim,* Abraham J. Domb, and Amiram Goldblum*

Cite This: *ACS Omega* 2024, 9, 42709−42720

Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: High energy materials (HEMs) play pivotal roles in diverse military and civil-commercial sectors, leveraging their substantial energy generation. Integrating machine learning (ML) into HEM research can expedite the discovery of high-energy compounds, complementing or replacing traditional experimental approaches. This manuscript presents an application of our in-house Iterative Stochastic Elimination (ISE) algorithm to identify HEMs. ISE is a generic algorithm that produces reasonable solutions for highly complex combinatorial problems. In molecular discovery, ISE focuses on physicochemical properties to distinguish between different classes of molecules. Due to its long track record in discovering novel, highly active biomolecules, we decided to apply ISE to another type of molecular discovery: High-energy materials. Two distinct ISE models, Model A (92 HEMs) and Model B (169 HEMs), integrated non-HEMs for comprehensive analysis. The results showcase significant achievements for both Models A and B. Model A identified 69% of active molecules in Model B, of which 62% had the highest score. Model B identified 80% of active molecules in Model A, with 61% having the highest score among those 80%. Subsequently, Model C was developed, merging all active molecules (261) from Models A and B. Statistical data indicate that Model C is a high-quality model. It was used to screen and score nearly 2 million molecules from the Enamine database. We find 66 molecules with the highest score of 0.89, plus 8 with that score which are active molecules included in the learning set of Model C. From the 66 molecules, 21 (32%) contain at least one nitro group. In conclusion, this study positions the ISE algorithm as a potential tool for discovering novel HEM candidates, offering a promising pathway for efficient and sustainable advancements in high-energy materials research.



1. INTRODUCTION

High Energy Materials (HEMs) are a category of chemical compounds or mixtures known for their ability to rapidly release large amounts of energy under specific conditions. These conditions typically involve exposure to external stimuli such as heat, shock, friction, or ignition. When triggered, HEMs undergo a rapid decomposition process, often accompanied by an exothermic reaction, meaning heat is released during the reaction. This decomposition generates expanding gases and releases substantial energy, primarily as heat. The abrupt release of energy creates a shock wave and further accelerates the gas expansion., which collectively contribute to the destructive force of an explosion. HEMs hold significant importance across military and civil-commercial sectors due to their ability to release substantial energy upon activation. In military applications, HEMs are extensively employed in weaponry such as missiles, rockets, and artillery while simultaneously playing roles in the civilian realm, notably in the aviation industry and mining operations.[1−7]

The methods used in synthesizing, identifying, and characterizing HEMs, specifically through traditional chemical approaches, demand substantial investment in resources, pose risks to researcher's safety, and raise environmental concerns due to their unsustainable nature. Additionally, the inefficiencies of trial and error methods in discovering new HEMs have resulted in a limited success rate through traditional chemical means in recent years, highlighting the need for predictive tools in this field.[8,9]

Recognizing the necessity of improving the efficiency of material discovery has led to integrating theory, computation, and experimentation in a holistic approach, expediting material innovation. Over the past decade, data-driven methods like machine learning have become pivotal in the materials research toolkit.[8−11] .

Traditional approaches, such as those relying on chemical theory and calculations like density functional theory (DFT), often encounter obstacles like high computational costs and lengthy estimation cycles.[12] Machine learning (ML) has emerged as a promising alternative, particularly in predicting properties such as density, explosion velocity, and sensitivity.

Received: February 2, 2024
Revised: July 17, 2024
Accepted: July 19, 2024
Published: October 11, 2024

Notable methods include Quantitative Structure−Activity Relationship (QSAR) and techniques like neural networks and random forests. However, utilizing these methods presents several challenges, including the complexity and diversity of HEM properties necessitating robust algorithms, interpretability and transparency issues with complex models, and significant computational resources required for training complex models and, consequently, time.[12−14] In materials science there is a strong emphasis on the interpretability and simplicity of ML models. Simple, interpretable models mitigate the risk of overfitting and can be computationally less expensive, which is crucial in applications like ML for industrial and applied purposes.[15]

The introduction of the ″Iterative Stochastic Elimination″ (ISE) algorithm, an award-winning algorithm developed in our laboratory,[16,17] marks a promising advancement in high-energy materials research. This algorithm was purposefully designed to find optimal solutions for complex combinatorial problems.[18,19] Over recent years, we have successfully adapted this algorithm to identify potential drug candidates.[20−26]

In this study, we aim to assess the applicability of the ISE algorithm in identifying new high-energy material candidates. Our findings highlight the ISE algorithm's potential as a valuable tool for identifying novel candidates in the high-energy materials domain.

2. METHODS

2.1. Iterative Stochastic Elimination (ISE). Iterative Stochastic Elimination (ISE) has a long track record in providing solutions to problems in biomolecules and, in particular, molecular discovery.[18] ISE constructs classification models that identify differences in properties between a group of known active molecules and inactive (or less active) ones. Such models are subsequently used to screen and score many molecules to identify candidates for the desired activity.

2.1.1. Data Curation. This initial stage involves gathering a data set comprising no fewer than several dozen known active molecules, ideally showcasing diverse characteristics. These active molecules are sourced from specific Web sites, scientific literature, and patents. Initially, mixtures and ionic compounds are removed from the data set, focusing solely on molecular compounds. Subsequently, the molecular data undergoes transformation into the Simplified Molecular Input Line Entry System (SMILES) format. Attention is dedicated to rectifying any discrepancies, errors, or extraneous elements present within the data set. The overarching objective of this curation process is laying a foundation for subsequent computational analyses.[27,28]

2.1.2. Tanimoto Index. To mitigate the potential for bias arising from molecular similarity, we utilize the Tanimoto Index (TI), also known as the Tanimoto coefficient or Jaccard similarity coefficient. It is defined by equation 1, where ″a″ and ″b″ represent the counts of binary features set to 1 in the fingerprints of compounds A and B, respectively, and ″c″ signifies the count of shared features present in both fingerprints.

$$\mathrm{TI}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{c}{a + b - c} \quad (1)$$

This index's value ranges from 0 to 1, indicating the degree of resemblance between molecules, with 0 signifying no shared features and one denoting complete identity.[29,30]

One of the tools enabling the calculation of the Tanimoto Index is the Chemistry Development Kit (CDK) nodes[31] integrated within the KNIME (version 2.10) platform.[32]

2.1.3. 2D MOE Descriptors. Molecular activity is a result of molecular properties ("descriptors"). However, thousands of molecular descriptors have been reported in the literature, having been computed using various methods.[33] Among these, we focused on 206 2D descriptors calculated using the MOE software. These descriptors quantitatively represent the physicochemical features of molecules, such as charge distribution, volume, molecular graphs, surface area, and more.[34] The descriptors are calculated for each molecule in the learning set (actives and decoys).

2.1.4. Applicability Domain and Picking Decoys. The random molecules ("decoys") are chosen from a large "molecular space" which is ideally used for later screening. To prevent "comparing apples with oranges" we use AD to ensure that decoys and active compounds share a few similar properties, thereby confining the selection of random molecules for the learning set. We calculate the average value and standard deviation among the actives for each property used to limit the choice of randoms. We then randomly select molecules whose properties fall within the range of the standard deviation of the actives. This approach increases the complexity of searching for differences in properties between the two classes, assuming the selected molecules are inactive. Consequently, this process adds a layer of complexity to the classification task.[35,36]

Determining the optimal ratio between decoys (random molecules) and active molecules involves considering several key factors, mainly determined by balancing the need for a robust model (suggesting picking more decoys) and computation time (more decoys means longer computations). When the primary objective is to identify active molecules from a large pool of candidates, it is often beneficial to use a high ratio of decoys to active compounds. Conversely, in situations where resources are limited and a more focused approach is needed, a low ratio of decoys to active compounds may be preferred. In such cases, the number of decoys may be similar to the number of active molecules. This approach prioritizes cost-effectiveness and efficient resource allocation. Another factor that may necessitate a lower ratio of decoys compounds to active is the selection of decoys based on the model's Applicability Domain (AD). When more AD descriptors are considered, it often results in a smaller pool of potential decoys, naturally leading to a reduced ratio between the number of decoys and active molecules.[37−39]

In our lab, that focuses mainly on discovering new drug candidates, it is common practice to select a ratio of 100 decoys for each active molecule. This approach closely reflects real-world scenarios, addressing the probability of 'encountering' active molecules within extensive chemical databases.[21,40] We also found that an AD of four properties (logP, Molecular Weight, H-bond donors and H-bond acceptors) is sufficient for discovering novel and highly efficient candidates.

2.1.5. Learning Set Construction. This process is performed using KNIME (version 2.10).[32] The n active molecules are combined with the decoys to create the learning set, comprising 101n molecules, each characterized by 206 2D descriptors. Descriptors with low variance across the set or highly correlated with other descriptors (having a Pearson correlation coefficient greater than 0.9) are excluded. Figure 1 illustrates the various steps for obtaining the learning set.

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

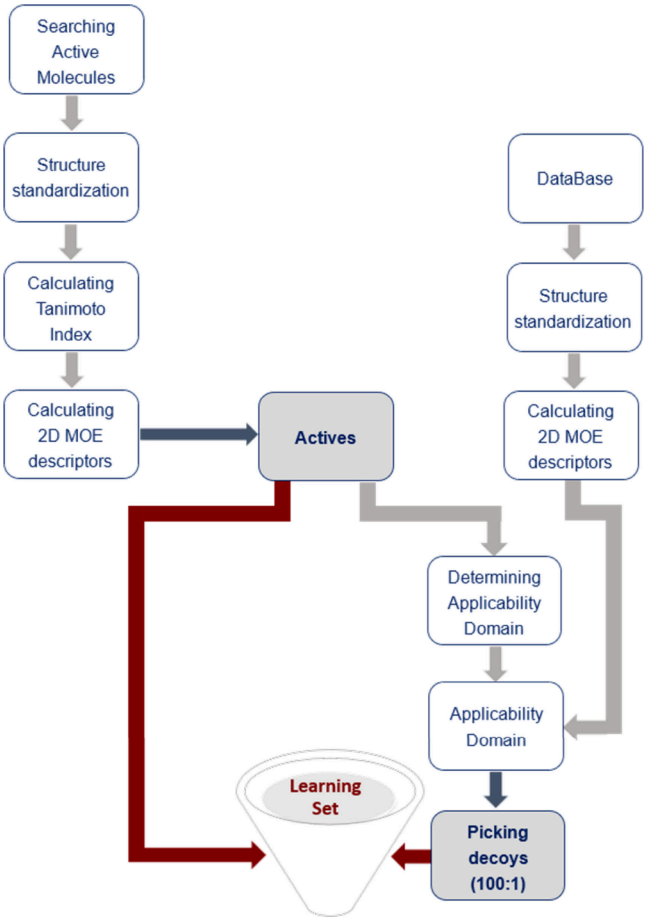


Figure 1. Stages leading to the formation of the learning set

2.1.6. Selecting and Evaluating Filters. After removing descriptors with low variance or high correlation, approximately 150−200 properties remain. The data set is divided into 5 "folds," each containing about 20% of actives and decoys. Four folds produce a model, while the model screens and scores the fifth. This process is repeated five times to screen and score the entire learning set.

To create a model, we randomly select five properties to form a "filter" through which we screen the data set. The total number of possible combinations is given by C = N!/[5!(N-5)!], so that if N = 200, the total number of possibilities would be ∼ 2.5 billion. However, we avoid computing that complete set and focus only on the effect of properties on the ability of filters to distinguish between the actives and decoys. We randomly select the 5-property filters, each with a score that reflects its classification ability. Filters are picked multiple times, ensuring each property is represented for analyzing its effects on classification ability. All the learning set molecules (for which we know the activity) are screened by each filter to determine its Matthews correlation coefficient (MCC, Equation 2)

Figure 2 illustrates the process of screening molecules through a single filter constructed from five distinct ranges of physio-chemical descriptors. Compound A successfully passes through all five ranges, yielding a positive index value. Depending on its actual status, this molecule can be classified as a TP if it is an active compound or an FP if it is a decoy. In contrast, compound B falls short of satisfying the criteria defined by all five ranges, resulting in a negative index value for

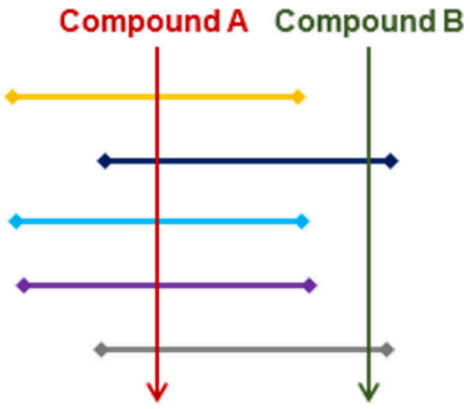


Figure 2. Screening molecules through a single filter, composed of ranges of five different physicochemical properties ("descriptors", horizontal lines)

that specific molecule. In terms of classification, compound B could be considered a TN if it is a decoy or an FN if it is an active compound:

Properties consistently associated with the poorest MCC values are systematically removed, resulting in subsequent iterations with a smaller set of properties. As iterations progress, the number of combinations decreases until the total number of possible combinations falls below one million, enabling an exhaustive search of all remaining combinations within a reasonable computational time frame.

All filters are produced and ranked based on their MCC scores, from the top MCCs to the lowest. Filters generated through the five folds are merged, and similar filters are eliminated to prevent bias in the final model.[18,41]

2.1.7. Screening Databases. In the information age, researchers use vast databases to link molecular structures with properties and forecast attributes of unseen molecules.[42] Following model validation, the ISE model may screen any molecular database to score molecules for which physicochemical properties have been assigned. We usually screen huge libraries with millions of molecules to score them by the model and pick the top-scored ones for experimental validation.

2.2. Statistical Measurement. Throughout model construction, several criteria allowed us to evaluate the model. This section presents each of these criteria used for assessing the qualities of our models.

2.2.1. MCC. The MCC (Matthews Correlation Coefficient) serves as a metric to assess the efficacy of classification processes, measuring their ability to distinguish between different groups. Each filter is evaluated based on its ability to differentiate between the active entities and the decoys based on Equation 2:

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP}) \times (\mathrm{TP} + \mathrm{FN}) \times (\mathrm{TN} + \mathrm{FP}) \times (\mathrm{TN} + \mathrm{FN})}} \quad (2)$$

TP corresponds to an active molecule that successfully satisfies all of the filter's criteria, TN represents a decoy molecule that fails to meet at least one of the filter's criteria, FP refers to a decoy molecule that successfully satisfies all of the filter's criteria, and FN indicates an active molecule that fails to meet at least one of the filter's criteria.

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

Every filter is assigned an MCC score within the range between the worst (−1) and the best (+1): a value of (0) indicates that the prediction's accuracy is comparable to a random choice.[43,44]

The MCC measure of success in classification is helpful for unbalanced sets, such as the learning set produced in this research, with a 100:1 ratio of decoys to actives.[45,46]

2.2.2. Precision, Accuracy and Recall. In classification models, Precision, Accuracy, and Recall are key metrics collectively assessing the model's accuracy in identifying positive instances, overall correctness, and proficiency in capturing all true positives, respectively.

Precision measures the model's ability to accurately identify positives in the data set (Equation 3):

$$\text{Precision} = \frac{\text{TP}}{\text{P}} = \frac{\text{TP}}{(\text{TP} + \text{FP})} \tag{3}$$

Precision values range from 0 to 1, with 1 signifying all positive identifications are correct and 0 indicating all are incorrect.[47,48]

Accuracy measures the proportion of correctly predicted instances out of the total cases (Equation 4), but its utility is limited in highly unbalanced learning scenarios.[49]

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}} \tag{4}$$

Recall (also known as TPR, True Positive Rate) evaluates the model's proficiency in identifying active molecules, calculated as the ratio of true positives to the sum of true positives and false negatives (FN) (Equation 5).

$$\text{Recall} = \text{TPR} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

Recall values range from 0 to 1, with 1 indicating the identification of all true positives and 0 denoting the failure to identify any true positives.[47,48]

2.2.3. Enrichment Factor (EF). The EF is employed to evaluate the benefits of virtual screening over random selection. It measures the fraction of correctly assigned true positives relative to the chance of finding true positives, which is the number of active molecules A divided by the total number N of molecules screened. The EF value is determined by the Equation 6a:

$$\text{EF} = \frac{\text{TP}/(\text{TP} + \text{FP})}{(\text{TP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})} = \frac{\text{TP}/(\text{TP} + \text{FP})}{\text{A/N}} \tag{6a}$$

In our ISE models, with a typical selection of decoys of 100:1, equation 6a may be presented as 6b:

$$\text{EF} = \frac{\text{TP}/(\text{TP} + \text{FP})}{(\text{TP} + \text{FN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})} = \frac{\text{Precision}}{1/101} = 101 \times \text{Precision} \tag{6b}$$

The resulting EF values from this equation range between 1 and > 100, where a value of 1 denotes randomness, while higher values suggest increased "enrichment".[50,51]

2.2.4. ROC Curve and AUC-ROC. The ROC curve is a fundamental evaluation tool for machine learning models, illustrating the relationship between TPR and False Positive Rate (FPR) across various classification thresholds. Each point on the curve corresponds to a specific classification threshold, representing a pair of TPR and FPR values. TPR, derived from Equation 5, represents the accurately identified true positive molecules, portraying the active fraction.

$$\text{FPR} = \frac{\text{FP}}{\text{FP} + \text{TN}} = \frac{\text{FP}}{\text{N(decoys)}} \tag{7}$$

Conversely, FPR, calculated using Equation 7, indicates the true negatives erroneously labeled as positive, representing the decoy fraction.

Figure 3 represents the ROC curve, where achieving a TPR of 100% and an FPR of 0% (at coordinates (1,1) and (0,0), respectively) indicates an instance of perfect classification. The Area Under the ROC Curve (AUC-ROC) quantifies model performance, ranging from 0 (indicating an inverted classifier) to 1 (representing a perfect classifier), with values around 0.5 suggesting randomness.[52−54]

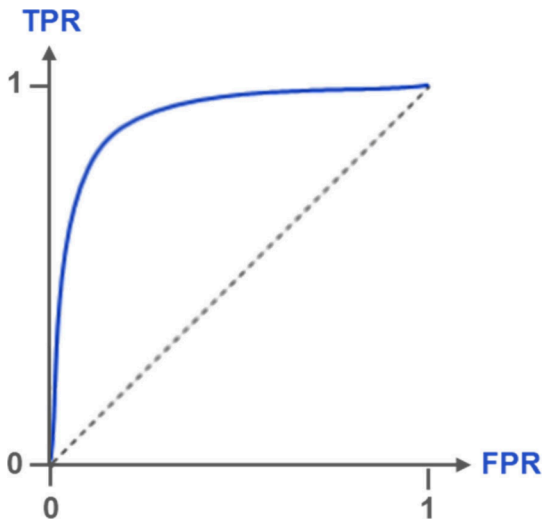


Figure 3. ROC curve.

2.2.5. Permutation Test (Y-Randomizations). The Permutation Test is a statistical method used to validate the stability of a model's statistical parameters to eliminate chance success. This test evaluates whether the model's classification ability surpasses random outcomes by comparing the model's performance on both original and shuffled data. It considers the model's predictive strength against random expectations by generating a distribution of performance metrics through multiple recalculations of model parameters like AUC-ROC values on permuted data. The resulting p-value is then computed as the proportion of iterations in which the model's performance on shuffled data exceeds that of the corresponding observed data. A low p-value indicates that the model's performance significantly surpasses what would be expected by random chance.[55,56]

To assess the statistical significance of our models, we conducted a Permutation Test using a Python script developed by ChatGPT for both Model A and Model B. In Model A, the initial data set comprised 92 instances labeled as '1' (actives) and 9200 instances labeled as '0' (decoys), while in Model B, there were 169 instances labeled as '1' and 16900 instances labeled as '0'. Each test involved 1000 iterations, wherein the

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

data set was shuffled while preserving the respective counts of '1' and '0' instances. The script calculated essential parameters for the Confusion Matrix (TP, FN, TN, and FP) throughout each iteration based on the permuted data. These values were the foundation for computing various performance metrics, encompassing TPR, FPR, AUC, Precision, Accuracy, and EF. The EF involves identifying the top-performing instances.

2.3. Molecular Index. The Molecular Index (MI) is employed to evaluate the predicted molecular activity of a molecule by scoring it based on filters. Each filter differentiates between positive (active) and negative (inactive or decoy) molecules. Filter effectiveness is quantified by values of true and false positives and negatives. MI calculation includes four metrics: TP, FP, TN, and FN. As shown in Equation 8, when computing the MI, two variables come into play: δ_{positive} and δ_{negative}. A δ_{positive} equals 1 if a molecule passes filter ″i″ as positive; otherwise, it is 0. Similarly, a δ_{negative} equals 1 if a molecule passes filter ″i″ as negative; otherwise, it is 0.

$$\text{MI} = \frac{\sum_{i=1}^{n} \delta_{\text{positive}_i} \frac{P}{P_f} - \delta_{\text{negative}_i} \frac{N}{N_f}}{n} \tag{8}$$

where P/P_f represents the ratio of true positives to false positives, and N/N_f represents the ratio of true negatives to false negatives for a given filter. P is true positives, P_f is false positives, N is true negatives, and N_f is false negatives. n is the total number of filters.

The resulting MI ranges from −1 to +1. A positive MI suggests activity, whereas a negative MI suggests inactivity. A higher MI implies an increased chance of experimental discovery as being active.[57]

All along the article, we use the terms ″index″ and ″score″ as being identical when discussing the values of MI.

3. RESULTS

From the outset of the study, our focus was solely on pure, uncharged HEMs. Therefore, in the quest for HEMs, we excluded mixtures and ionic compounds. Two separate models were constructed based on distinct data sources to assess whether the ISE algorithm is suitable for detecting potential HEMs.

3.1. Constructing the Data Sets. Model A incorporated HEMs obtained from various sources: an article on predicting the power of energetic materials,[58] the Explosives Database,[59] and the AIST Database.[60] For Model B, the active molecules were extracted from an article that included 222 HEMs.[61]

In each model, we implemented data curation as described in Methods. Also, to avoid potential bias due to molecular similarity, we computed the Tanimoto Index (TI). In each model, out of a pair of HEMs with a TI exceeding 0.9 we retained the one with a lower total TI in comparison to all other molecules. Additionally, we excluded molecules that appear in both Model A and Model B, retaining only one instance of each unique molecule. Finally, Model A comprised 92 HEMs, while Model B retained 169 HEMs, constituting a distinct set from the 92 molecules present in Model A.

A total of 206 physicochemical properties (2D descriptors) were computed for each active molecule (HEM) using MOE2018. Subsequently, the Applicability Domain (AD) was independently determined for Model A and Model B based on their respective values of the 206 2D descriptors. The selection of the 16 descriptors that define the AD was guided by their alignment within the range of plus or minus one standard deviation from the mean and their established relevance to high-energy materials.[54,58] The literature addressing energetic material properties emphasizes the influence of atomic composition and molecular weight within a compound's formula. Additionally, we extended our criteria to encompass various types of bonds. Model A and Model B have an identical number of 16 descriptors for applicability domain, each displaying distinctive value ranges within its respective model (Table 1). For instance, in Model A, the 'b_heavy' descriptor, signifying the number of bonds between heavy atoms (atoms with $z > 1$), ranged from 8 to 20, while in Model B, this range extended from 5 to 23. Similarly, the 'b_rotN' descriptor, indicating the number of rotatable bonds, exhibited ranges from 0 to 8 in Model A and 1 to 6 in Model B.

Table 1. Applicability Domain Descriptors for All Models

Atomic composition	Bond composition	Molecular weight
a_count	b_1rotN	MW
a_heavy	b_1rotR	
a_nC	b_ar	
a_nH	b_count	
a_nN	b_double	
a_nO	b_heavy	
	b_rotN	
	b_rotR	
	b_1rotN	

Table S1 specifies the code, description, and range of each of the 16 descriptors for the ADs of Model A and Model B.. Table S2 includes the Complete list of 2D descriptors calculated by MOE2018.

3.2. Creating and Evaluating the Models. Two ISE models were created using active molecules identified during preliminary stages and decoy molecules from the Enamine Database, which includes approximately 2 million compounds.[62] Similar to our approach with the HEMs, we conducted structure standardization and computed 206 2D descriptors for all molecules in the database.

For each model, the selection of decoy molecules from the Enamine database was limited to those within the relevant AD specific to that model. The ratio between active molecules and decoys was maintained at 1:100 in both models. As a result, the learning set of Model A consisted of 92 active molecules and 9200 decoys (Table S3), while the learning set of Model B comprised 169 active molecules and 16900 decoys (Table S4).

Table 2 presents the MCC minima, maxima, and mean values and the AUC-ROC scores for the five folds in each model.

In Figure 4, scatter plots are presented for Models A and B, illustrating the scoring outcome of each molecule using the complete set of model filters. In both cases, the red dots represent the decoy molecules (random), while the blue diamonds represent the HEMs.

Table S5 presents the values of the Confusion Matrix (CM) at various cutoff indexes for Model A and Model B, respectively. These tables also include the computed results of TPR (using Equation 5), FPR (using Equation 7), and EF (using Equation 6b) based on those values.

Table 3 presents additional parameters derived from the CM values, offering valuable insights into model performance. These parameters include the calculation of MCC (using

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

Table 2. Statistical Data for Models A and B across Five Folds

	Model A				Model B			
	MCC				MCC			
	min	max	mean	AUC-ROC	min	max	mean	AUC-ROC
Fold 1	0.86	0.87	0.86	0.97	0.81	0.86	0.82	0.97
Fold 2	0.83	0.86	0.84	0.90	0.81	0.85	0.82	0.99
Fold 3	0.82	0.84	0.82	0.99	0.81	0.85	0.82	0.99
Fold 4	0.80	0.84	0.81	0.98	0.80	0.85	0.81	1.0
Fold 5	0.86	0.86	0.86	0.99	0.82	0.85	0.83	0.99

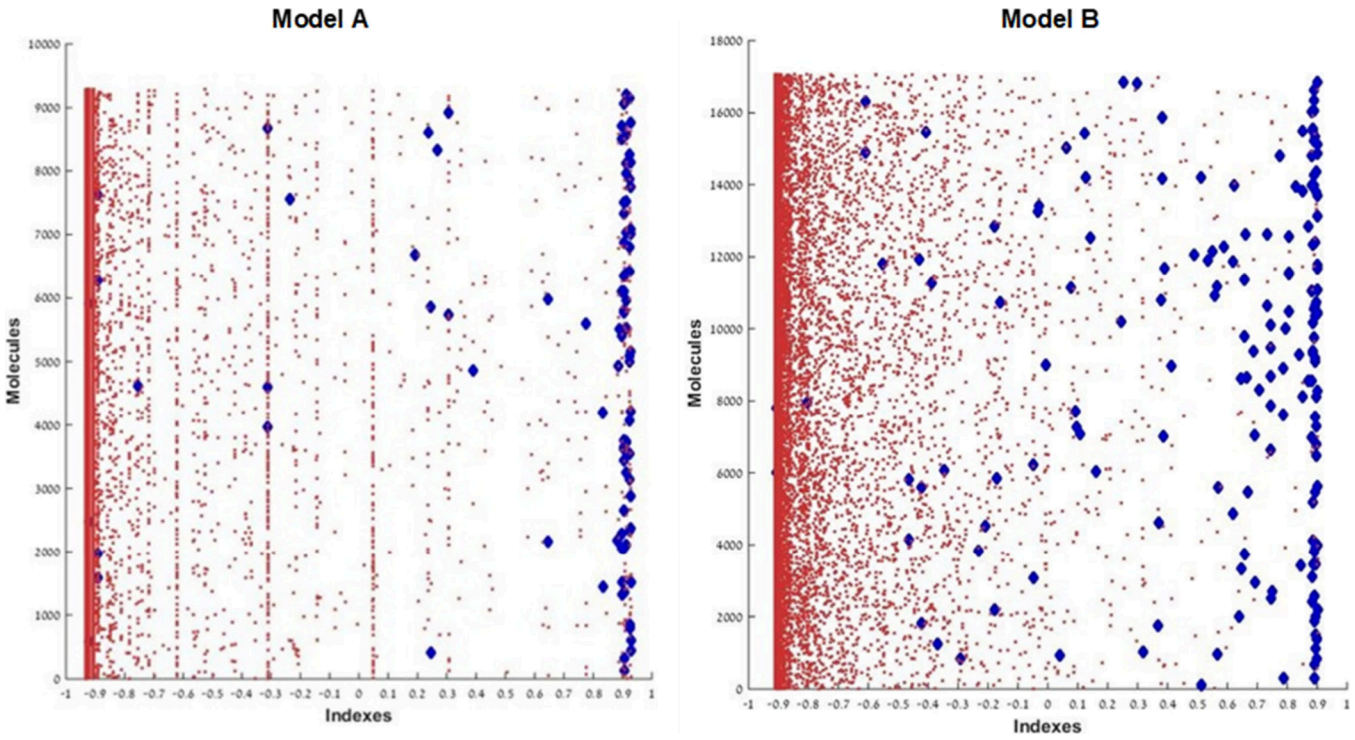


Figure 4. Scatter Plots of Models A and B HEMs in blue diamonds and decoys in red dots.

Table 3. Evaluation Metrics for Models A and B

	Model A			Model B		
	MCC	EF	TP/FP	MCC	EF	TP/FP
Index 0	0.35	15	0.18	0.53	34	0.51
Index 0.1	0.41	20	0.25	0.55	39	0.62
Index 0.2	0.41	21	0.26	0.57	43	0.74
Index 0.3	0.41	22	0.27	0.61	49	0.96
Index 0.4	0.43	25	0.32	0.63	57	1.30
Index 0.5	0.44	26	0.35	0.65	62	1.59
Index 0.6	0.45	27	0.36	0.65	67	1.96
Index 0.7	0.46	29	0.40	0.62	71	2.32
Index 0.8	0.48	32	0.46	0.58	73	2.67
Index 0.9	0.48	34	0.50	0.24	71	2.33

Equation 2), as well as EF (equation 6a,6b) and the TP/FP ratio (see discussion) above each index (score).

The ROC curves for Models A and B (Figure 5) were constructed using each index's TPR and FPR values. Subsequently, the AUC-ROC value was calculated for each model based on the generated curves.

Cross-Validation of the Models. Models A and B, including 92 and 169 HEMs, respectively, were employed for mutual screening. In this process, the 169 active molecules of

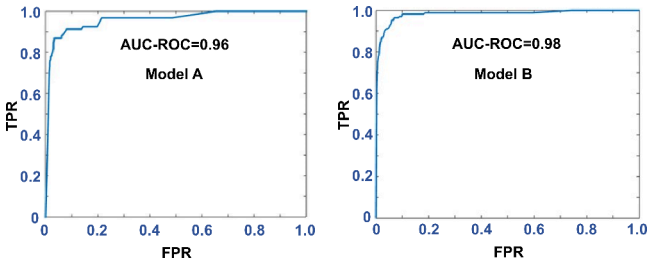


Figure 5. ROC Curves for Models A and B.

model B were screened and scored using the Model A filters, and vice versa. There is no overlap in active molecules between Models A and B. Random molecules were selected from a pool of nearly two million, ensuring a meager chance of random similarity. The screening results of both models are presented in Table 4.

4. DISCUSSION

Our in-house algorithm (ISE) has been successfully utilized for over a decade for detecting novel candidate drugs. However, molecules can be directed to different purposes, and HEM is just another challenge for molecular discovery, given their

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

Table 4. Cross Screening Results

	Model A[a]	Model B[b]
No. of molecules screened	169	92
No. of hits[c]	117	74
No. of hits with the highest index (>0.9)	73	45

[a]Model B HEMs by Model A filters (2nd column). [b]Model A HEMs by Model B filters (3rd column). [c]Any molecule with a cutoff of > 00 is defined as a "hit".

primary shared basis with drugs, which is physicochemical properties. In both types of activities, we need previous data to produce new data, and the significant difference is that drugs are aimed at biological targets while HEMS are not.

To evaluate the ISE model's capability in identifying new candidates for high-energy materials, we constructed two separate ISE models for HEMs, explored their ability through cross-screening between them and found it to be of very good quality. These models consist of filters designed to distinguish between known actives (HEMs) and randomly picked molecules (non-HEMs), as explained in the Methods section. This process closely resembles the application of ISE for drug discovery, involving the construction of learning sets heavily "diluted" with inactive molecules.

4.1. Validation of the Models. Table 2 outlines the statistical metrics for Model A and Model B across five separate folds. The results suggest that both Model A and Model B exhibit strong predictive abilities. Model A's MCC values range from 0.80 to 0.87, with a mean of 0.84 across the five folds. This model consistently demonstrated strong discrimination ability, with AUC-ROC scores ranging from 0.90 to 0.99. Model B also exhibits similar MCC values between 0.80 and 0.86 (mean = 0.82), and Its AUC-ROC ranges from 0.97 to 1.0. Notably, both models achieved high AUC-ROC scores, indicating their ability to differentiate between active and inactive molecules effectively. The results from both Model A and Model B emphasize the efficiency of ISE in detecting the differences between the two classes. The consistently high MCC and AUC-ROC scores affirm the suitability of these models for predictive screening purposes.

4.2. Major Descriptor Families and Their Prevalence. Model A consists of 442 filters, each comprising five physicochemical properties. In contrast, Model B included 2887 filters. Due to the large number of filters, we limit the number to the top 20% of filters (577 filters) for the cross-screening (see section 3).

We analyzed the appearance of each descriptor in both models and subsequently organized them into distinct types (categories) based on the MOE assignments. Among these categories, we focused on those containing at least 10% of the total descriptors in a given model. We found that some types of descriptors are identical in both Model A and Model B. These shared categories include 'Atom Counts and Bond Counts', 'Pharmacophore Features', and 'Physical Properties'. Notably, these three descriptor types constitute the majority of descriptors in both models, contributing to 80% of the descriptor distribution in Model A and a more significant 95% in Model B. A graphical representation of the distribution of each type in both models is presented in Figure 6.

Regarding descriptor prevalence, the 'Atom Counts and Bond Counts' type exhibits the highest frequency in both Model A and Model B, comprising 53% and 36% of the descriptors, respectively. On the other hand, descriptors from the 'Physical Properties' type contribute only 11% and 27% in Model A and Model B, respectively. Detailed descriptors for both models are provided in Table S6.

4.3. Scatter Plots. It is evident from both scatter plots (Figures 4A and 4B) that the models effectively distinguish between HEMs and decoy molecules. Approximately 75% of the active molecules passed through the filters with a score of ≥ 0.7, while $\sim$ 85% of the randoms (assumed to be inactive molecules) failed to pass the filters (Index ≤ -0.8).

4.4. Evaluating the Models. The performance of the ISE models in distinguishing active from inactive molecules is shown in Table 3. Metrics such as MCC, EF, and TP/FP were evaluated across index thresholds for Models A and B. Model A maintains a relatively stable MCC range of 0.35 to 0.48, reflecting consistent discrimination ability. EF for Model A ranges from 15 to 34, while TP/FP varies from 0.18 to 0.50. In contrast, Model B exhibits larger fluctuations in MCC (0.24 to 0.65) and associated EF (34 to 73), with varying TP/FP (0.51 to 2.67).

Model A shows consistent performance in MCC across the index thresholds, indicating reliability in predictions across different parameters. While Model B demonstrates higher EF

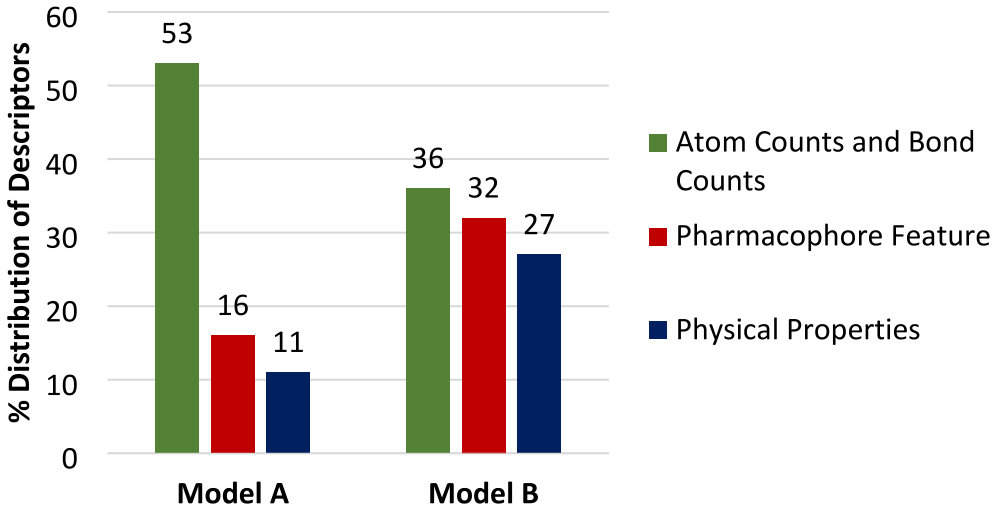


Figure 6. Major descriptor types for Models A and B.

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

values, its MCC values show greater variability, suggesting potential inconsistency in predictive accuracy. Model B's TP/FP ratios imply improved discrimination between true and false positives.

4.5. ROC Curves. The ROC curve analysis for Models A and B (Figure 5) yields a similar Area Under the Curve (AUC) of 0.96 (Model A) and 0.98 (Model B). These high AUC values underscore both models' consistent and robust discriminatory capacity in effectively distinguishing between positive and negative classes.

4.6. Y Randomization Test. Conducting 1000 iterations in this test allows us to evaluate the model's performance against a randomized baseline. For Model A, the computed mean AUC value is entirely random, 0.50. Similar indications are given by the TPR and FPR (~0.01 in both). Further details and visual representations of the distributions for the performance metrics of Model A, including TPR, FPR, accuracy, precision, and AUC, can be found in the Supplemental section (Figure S1).

The Y-Randomization results in Model B closely mirrored those obtained for Model A. Both models exhibited close alignment in outcomes, with results matching up to two decimal places. This consistency between Models A and B reinforces the reliability and robustness of our methodology across diverse models.

4.7. Cross Testing. The filters of each model were used to screen the active molecules (known HEMs) in the other model and received a score depending on whether they passed through the filter or failed. A molecule receiving a positive index indicates that the model has recognized it as HEM.

As evident from Table 4, Model A successfully identifies 69% of Model B's active molecules, with 62% of those receiving the highest index value of 0.91. In parallel, Model B identifies 80% of the active molecules of Model A, and 61% of those pass Model B's filters with a maximum score of 0.91. These outcomes underscore the model's competence in discerning active molecules from the respective data sets. The high percentage of accurately identified HEMs and hits, combined with the consistently elevated index values, reinforces the predictive power of both models. The findings further highlight the effectiveness of the applied screening approach in identifying high-energy molecules.

The graphical representation of the screening results in Figure 7 complements the quantitative data, providing a visual summary of the models' discriminative performance. Together, the tabulated results and visual depiction offer a comprehensive understanding of the model's capabilities in molecular screening and contribute to discussing their potential application in practical scenarios.

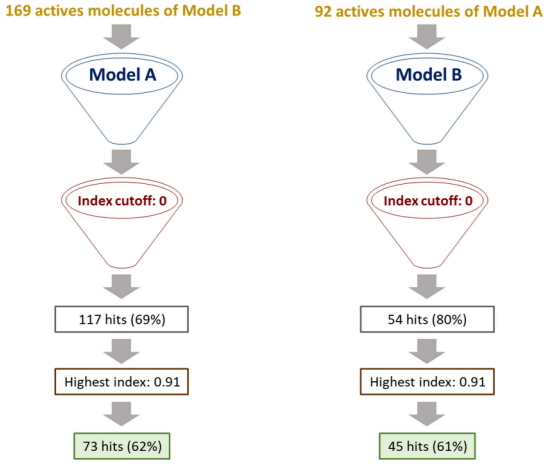


Figure 7. Models A and B screening results.

The effectiveness of Models A and B in molecular discrimination is visually depicted through histograms (Figure 8). The histograms provide a snapshot of each model's True Positive Rate (%TPR) across varying decision cutoffs. Model A's histogram showcases a progressive decline in %TPR as the cutoff values increase. The highest %TPR, 69%, is observed at the lowest cutoff of 0. In contrast, Model B's histogram demonstrates consistent %TPR values, ranging from 76% to 60%, for cutoffs between 0.3 and 0.7. Notably, at cutoffs of 0.5 and 0.6, Model B's %TPR drops to 60%.

Analyzing these histograms provides valuable insights into the behavior of the models at varying levels of stringency. Model A demonstrates higher %TPR values for lower cutoffs, indicating its ability to identify active molecules even when using more permissive thresholds. In contrast, Model B maintains relatively stable %TPR values across intermediate cutoffs, suggesting consistent performance in that range.

The results obtained from this research confirm the ability of ISE models to identify diverse, high-energy candidates.

4.8. Nitro Groups and AD. While it is true that Enamine was not originally designed to include explosives rich in nitro groups, we would like to provide additional data regarding the presence of molecules with nitro groups in our data set. Among the 1938610 molecules in the Enamine database, 43,425 molecules (2.24%) contain nitro groups. Due to random picking of decoys, we expect to pick 206 (2.24%) molecules with nitro groups out of the 9,200 decoys for Model A. For Model B, we expect to pick 379 molecules with nitro groups out of the 16,900 decoys. However, we find many more due to our applicability domain: 449 (4.88%) decoys with nitro groups for model A and 684 (4.05%) decoys with nitro groups in model B. This result supports our decision for the properties used for AD, nitro group not being among them.

4.9. Constructing a larger model. Given that the results from cross-screening Model A and Model B confirmed the ability of ISE models to identify high-energy candidates, we proceeded to construct a third model, Model C. We combined the HEMs from Models A and B in order to have more HEMS in a single model and to examine whether that model could have better performance that the two smaller data sets. Model C, thus comprises 261 active molecules from Models A and B. The construction steps of Model C, including the AD and the picking of decoys for the learning set, are identical to those of Models A and B, as detailed in the Methods section.

According to Table 5, the results across all folds consistently demonstrate high

MCC values, indicating robust performance in binary classification tasks. Additionally, the AUC-ROC scores consistently near 0.99 signify excellent discriminatory ability of Model C. This suggests that Model C is effective in accurately classifying instances and that it may be used for screening external data sets.

The confusion matrix (CM, Table 6) illustrates a tradeoff between true positive (TP) and false positive (FP) rates with increasing index thresholds. In the higher thresholds the FP instances are reduced with respect to TP. All three criteria − MCC, EF and TP/FP increase with increasing index threshold. The top molecules in the model obtain a maximal index of

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

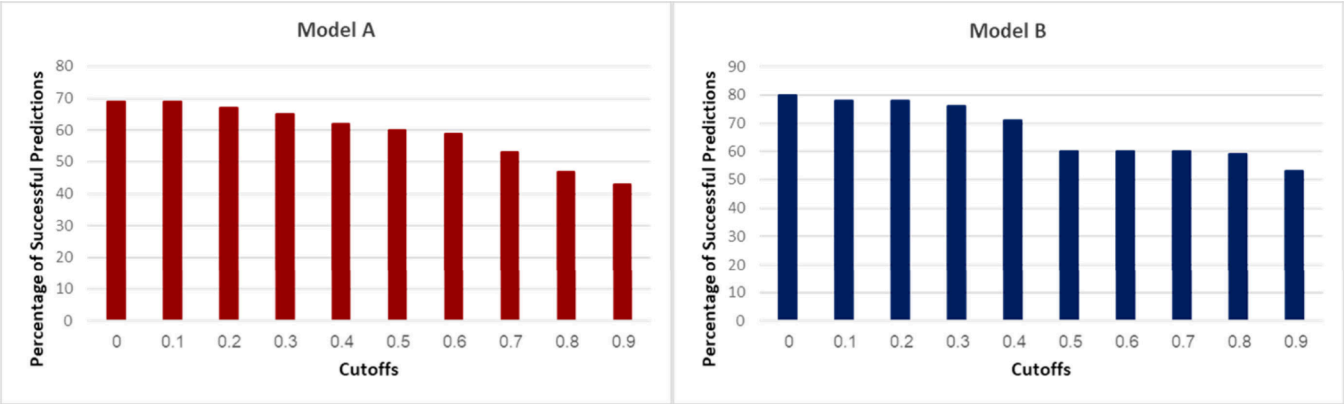


Figure 8. Histograms depicting the effectiveness of Models A and B in molecular discrimination.

Table 5. Statistical Data for Model C across Five Folds

	MCC			
	min	max	mean	AUC-ROC
Fold 1	0.79	0.81	0.80	0.97
Fold 2	0.80	0.81	0.80	0.99
Fold 3	0.79	0.82	0.80	0.99
Fold 4	0.79	0.82	0.80	0.99
Fold 5	0.77	0.82	0.78	0.99

0.89. For practical decisions (external screening) we pick the best index threshold with the largest TP/FP, reducing the total number of molecules to be tested.

4.10. Screening by Model C. Consequently, we screened 1938610 molecules from the Enamine Database by the filters of Model C. We find 3,156 molecules with a positive index (above 0.0). More than 800 (∼25% of the positives) of those contain at least one nitro group. Subsequently, upon applying the acceptable cutoff for further testing (0.7), we obtained 331 molecules, out of which 115 (∼35%) have at least one nitro group. The maximum index achieved in Model C is 0.89, with 74 molecules reaching this index, of which 8 are part of the active compounds within the learning set. Notably, among the remaining 66 molecules, 21 (32%) contain at least one nitro group. It is thus clear that with increasing scores we find a larger proportion of nitro-containing HEM candidates.

4.11. The Importance of TP/FP Values. TP/FP ratios are results of the model, and have different values depending on the index cutoff, as seen in Tables 3 and 6. Screening by the model produces a plethora of indexes for the huge number of molecules. We aim at picking the screened molecules with the largest indexes for testing by real experiments. That number should be limited by several factors: 1) the number of molecules that can be tested (50 molecules/week or 5 per week ?) and 2) the indexes of the top scoring molecules and their number. Assuming that by screening ∼ 2 million molecules by model C (Table 6) we find 20 molecules with index > 0.75. The TP/FP of the cutoff 0.7 (∼2) means that out each 3 molecules (TP/FP = 2/1) sent for testing, 2 may be found to be active based on the results of the model. However, sending only 3 molecules is risky and so we can decide to send 6 (with a chance of discovering 4 actives) or even send 18 (with the chance to discover 12 actives). As all our screened molecules are commercially available and without attached patents, we are only limited by prices and by the limits of the experimental throughput.

5. CONCLUSIONS

Before this study, the ISE algorithm had been successfully utilized to identify new drug candidates that were subsequently tested in laboratory settings. The primary aim of this current research was to investigate whether the ISE could be adapted to develop models capable of identifying potential energetic materials (HEMs). To achieve this objective, we constructed two ISE models, Models A and B. Through comprehensive evaluations utilizing statistical metrics such as MCC and AUC-ROC, both models demonstrated robust discriminatory abilities. In the later analyses involving cross-screening, scatter plots, ROC curves, and histograms, the model's effectiveness in distinguishing HEMs from decoys was confirmed.

In the subsequent phase, both models screened each other's active molecules, yielding impressive results. Model A

Table 6. Confusion Matrix Values of Model C

	TP	FP	FN	TN	MCC	EF	TP/FP
Index 0	209	585	52	25515	0.45	27	0.36
Index 0.1	198	364	63	25736	0.51	36	0.54
Index 0.2	189	268	72	25832	0.54	42	0.71
Index 0.3	187	226	74	25874	0.56	46	0.83
Index 0.4	180	173	81	25927	0.59	52	1.04
Index 0.5	172	100	89	26000	0.64	64	1.72
Index 0.6	162	83	99	26017	0.64	67	1.95
Index 0.7	155	77	106	26023	0.63	67	2.01
Index 0.8	153	58	108	26043	0.65	73	2.64
Index 0.9	0	0	261	26100	NaN	NaN	NaN
Index 1	0	0	261	26100	NaN	NaN	NaN

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

identified 89 out of 169 of Model B's active molecules as HEMs, with 82% achieving the highest index score (0.91). Similarly, Model B identified 54 out of 92 active molecules from Model A as HEMs, with 83% scoring the highest index of 0.91.

Based on the promising outcomes of Models A and B, a third model, Model C, was developed, which includes all 261 active molecules from Models A and B. Thanks to its robust performance in effectively distinguishing between active compounds and decoys, Model C was utilized to screen and score the Enamine database (~2 million molecules). Out of approximately 2 million molecules screened using Model C, 74 achieved the highest score of 0.89. Further analysis revealed that 8 of these top-scoring molecules were part of the 261 active molecules in Model C's learning set. For the remaining 66 molecules, we will determine which ones are not patented, and acquire them for calorimetric analysis, such as Differential Scanning Calorimetry (DSC).

■ ASSOCIATED CONTENT

Data Availability Statement

All relevant data are part of the Supporting Information. The ISE algorithm is described in the Methods section and was thoroughly detailed in a review by Stern and Goldblum.[18]

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.4c01070.

Applicability domains of Models A and B, complete list of 2D descriptors calculated by MOE 2018, active and decoy molecules used to construct the learning sets, values of the confusion matrix (CM) at different cutoff indexes for Models A and B, common type descriptors in Models A and B with respective prevalence percentages, visual representations of the distributions for the performance metrics obtained from the Y randomization test for Model A (XLSX)

■ AUTHOR INFORMATION

Corresponding Authors

Mazal Rachamim − *Molecular Modelling and Drug Design Lab, Institute for Drug Research and Fraunhofer Project Center for Drug Discovery and Delivery, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91905, Israel*; orcid.org/0000-0001-8093-832X; Email: Mazal.Rachamim@mail.huji.ac.il

Amiram Goldblum − *Molecular Modelling and Drug Design Lab, Institute for Drug Research and Fraunhofer Project Center for Drug Discovery and Delivery, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91905, Israel*; Email: amiramg@ekmd.huji.ac.il

Author

Abraham J. Domb − *The Institute for Drug Research, School of Pharmacy, Faculty of Medicine, The Hebrew University of Jerusalem, Jerusalem 91905, Israel*; orcid.org/0000-0002-2241-7726

Complete contact information is available at: https://pubs.acs.org/10.1021/acsomega.4c01070

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors have no acknowledgements to Declar

■ ABBREVIATION

AD	Applicability domain
AUC-ROC curve	Area under the receiver operating characteristics curve
CM	Confusion matrix
EF	Enrichment factor
FN	False-negative
FP	False positive
FPR	False positive rate
HEMs	High-energy materials
HTS	High-throughput screening
ISE	Iterative stochastic elimination
MCC	Matthews correlation coefficient
MI	Molecular index
ML	Machine learning
MOE	Molecular operating environment
PubChem	Public chemical database
ROC curve	Receiver operating characteristics
SMILES	Simplified molecular input line entry system
TI	Tanimoto index
TN	True negative
TP	True positive
TPR	True positive rate

■ REFERENCES

(1) Zapata, F.; García-Ruiz, C. Chemical classification of explosives. *Crit. Rev. Anal. Chem.* **2020**, *51* (7), 656−673.

(2) Klapötke, T. M. *Chemistry of High-Energy Materials*; De Gruyter, 2017.

(3) Beveridge, A. *Forensic Investigation of Explosions*; CRC Press, 1998.

(4) Agrawal, J. P. *High Energy Materials: Propellants, Explosives and Pyrotechnics*; John Wiley & Sons, 2010.

(5) Keshavarz, M. H., Klapötke, T. M. *Energetic Compounds*; De Gruyter, 2020.

(6) Muthurajan, H.; Ghee, H. Software development for the detonation product analysis of high energetic materials-Part I. *Cent. Eur. J. Energ. Mater.* **2008**, *5* (3−4), 19−35.

(7) Christopher, I. L.; Michalchuk, A. A.; Pulham, C. R.; Morrison, C. A. Towards Computational Screening for New Energetic Molecules: Calculation of Heat of Formation and Determination of Bond Strengths by Local Mode Analysis. *Frontiers in chemistry* **2021**, *9*, No. 726357.

(8) de Pablo, J. J.; Jackson, N. E.; Webb, M. A.; Chen, L.-Q.; Moore, J. E.; Morgan, D.; Jacobs, R.; Pollock, T.; Schlom, D. G.; Toberer, E. S.; et al. New frontiers for the materials genome initiative. *npj Comput. Mater.* **2019**, *5* (1), 41.

(9) Vasudevan, R. K.; Choudhary, K.; Mehta, A.; Smith, R.; Kusne, G.; Tavazza, F.; Vlcek, L.; Ziatdinov, M.; Kalinin, S. V.; Hattrick-Simpers, J. Materials science in the artificial intelligence age: high-throughput library generation, machine learning, and a pathway from correlations to the underpinning physics. *MRS Commun.* **2019**, *9* (3), 821−838.

(10) Morgan, D.; Jacobs, R. Opportunities and challenges for machine learning in materials science. *Annual Review of Materials Research* **2020**, *50*, 71−103.

(11) Kang, P.; Liu, Z.; Abou-Rachid, H.; Guo, H. Machine-learning assisted screening of energetic materials. *The Journal of Physical Chemistry A* **2020**, *124* (26), 5341−5351.

(12) Tian, X. L.; Song, S. W.; Chen, F.; Qi, X. J.; Wang, Y.; Zhang, Q. H. Machine learning-guided property prediction of energetic

materials: Recent advances, challenges, and perspectives. *Energetic Materials Frontiers* **2022**, *3* (3), 177−186.
(13) Fournier, M.; Vroland, C.; Megy, S.; Aguero, S.; Chemelle, J. A.; Defoort, B.; Jacob, G.; Terreux, R. In silico Genotoxicity Prediction by Similarity Search and Machine Learning Algorithm: Optimization and Validation of the Method for High Energetic Materials. *Propellants, Explos., Pyrotech.* **2023**, *48* (4), e202200259.
(14) Zang, X.; Zhou, X.; Bian, H.; Jin, W.; Pan, X.; Jiang, J.; Koroleva, M. Y.; Shen, R. Prediction and construction of energetic materials based on machine learning methods. *Molecules* **2023**, *28* (1), 322.
(15) Chen, C.; Zuo, Y.; Ye, W.; Li, X.; Deng, Z.; Ong, S. P. A critical review of machine learning of energy materials. *Adv. Energy Mater.* **2020**, *10* (8), 1903242.
(16) Symposium on Emerging Computational Technologies, Washington D.C., August 2000. http://oldwww.acscomp.org/Awards/emerging2000.html
(17) Algorithm dramatically improved drug discovery methods, earns 2017 Kaye Innovation Award, *EurekaAlert!*. June 22, 2017. https://www.eurekalert.org/news-releases/505913.
(18) Stern, N.; Goldblum, A. Iterative stochastic elimination for solving complex combinatorial problems in drug discovery. *Israel Journal of Chemistry* **2014**, *54* (8−9), 1338−1357.
(19) Zatsepin, M.; Mattes, A.; Rupp, S.; Finkelmeier, D.; Basu, A.; Burger-Kentischer, A.; Goldblum, A. Computational discovery and experimental confirmation of TLR9 receptor antagonist leads. *Journal of Chemical Information and Modeling* **2016**, *56* (9), 1835−1846.
(20) Cern, A.; Marcus, D.; Tropsha, A.; Barenholz, Y.; Goldblum, A. New drug candidates for liposomal delivery identified by computer modeling of liposomes' remote loading and leakage. *J. Controlled Release* **2017**, *252*, 18−27.
(21) Da'adoosh, B.; Marcus, D.; Rayan, A.; King, F.; Che, J.; Goldblum, A. Discovering highly selective and diverse PPAR-delta agonists by ligand based machine learning and structural modeling. *Sci. Rep.* **2019**, *9* (1), 1106.
(22) El-Atawneh, S.; Hirsch, S.; Hadar, R.; Tam, J.; Goldblum, A. Prediction and experimental confirmation of novel peripheral cannabinoid-1 receptor antagonists. *Journal of Chemical Information and Modeling* **2019**, *59* (9), 3996−4006.
(23) Da'adoosh, B.; Kaito, K.; Miyashita, K.; Sakaguchi, M.; Goldblum, A. Computational design of substrate selective inhibition. *PloS computational biology* **2020**, *16* (3), No. e1007713.
(24) El-Atawneh, S.; Goldblum, A. Candidate Therapeutics by Screening for Multitargeting Ligands: Combining the CB2 Receptor With CB1, PPARγ and 5-HT4 Receptors. *Frontiers in pharmacology* **2022**, *13*, 812745.
(25) Merk, H.; Amran-Gealia, T.; Finkelmeier, D.; Kohl, C.; Pichota, I.; Stern, N.; Rupp, S.; Goldblum, A.; Burger-Kentischer, A. Human-based immune responsive in vitro infection models for validation of novel TLR4 antagonists identified by computational discovery. *Microorganisms* **2022**, *10* (2), 243.
(26) Stepanenko, N.; Wolk, O.; Bianchi, E.; Wright, G. J.; Schachter-Safrai, N.; Makedonski, K.; Ouro, A.; Ben-Meir, A.; Buganim, Y.; Goldblum, A. In silico Docking Analysis for Blocking JUNO-IZUMO1 Interaction Identifies Two Small Molecules that Block in vitro Fertilization. *Front. Cell Dev. Biol.* **2022**, *10*, 824629.
(27) Hähnke, V. D.; Kim, S.; Bolton, E. E. PubChem chemical structure standardization. *J. Cheminform.* **2018**, *10*, 36.
(28) Fourches, D.; Muratov, E.; Tropsha, A. Trust, but verify: on the importance of chemical structure curation in cheminformatics and QSAR modeling research. *Journal of chemical information and modeling* **2010**, *50* (7), 1189.
(29) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug discovery today* **2006**, *11* (23−24), 1046−1053.
(30) Cereto-Massagué, A.; Ojeda, M. J.; Valls, C.; Mulero, M.; Garcia-Vallvé, S.; Pujadas, G. Molecular fingerprint similarity search in virtual screening. *Methods* **2015**, *71*, 58−63.
(31) Willighagen, E. L.; Mayfield, J. W.; Alvarsson, J.; Berg, A.; Carlsson, L.; Jeliazkova, N.; Kuhn, S.; Pluskal, T.; Rojas-Cherto, M.; Spjuth, O.; et al. The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching. *J. Cheminform.* **2017**, *9*, 33.
(32) Berthold, M. R.; Cebron, N.; Dill, F.; Gabriel, T. R.; Kotter, T.; Meinl, T.; Ohl, P.; Thiel, K.; Wiswedel, B. KNIME-the Konstanz information miner: version 2.0 and beyond. *SIGKDD Explor. Newsl.* **2009**, *11* (1), 26−31.
(33) Kamath, V.; Pai, A. Application of molecular descriptors in modern computational drug design-an overview. *Research Journal of Pharmacy and Technology* **2017**, *10* (9), 3237−3241.
(34) *Molecular Operating Environment (MOE)*, 2018.01; Chemical Computing Group ULC: Montreal, QC, Canada, 2018.
(35) Netzeva, T. I.; Worth, A. P.; Aldenberg, T.; Benigni, R.; Cronin, M. T.D.; Gramatica, P.; Jaworska, J. S.; Kahn, S.; Klopman, G.; Marchant, C. A.; et al. Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships: The report and recommendations of ecvam workshop 52. *Altern. Lab Anim.* **2005**, *33* (2), 155−173.
(36) Gadaleta, D.; Mangiatordi, G. F.; Catto, M.; Carotti, A.; Nicolotti, O. Applicability domain for QSAR models: where theory meets reality. *International journal of quantitative structure-property relationships (IJQSPR)* **2016**, *1* (1), 45−63.
(37) Tropsha, A. Best practices for QSAR model development, validation, and exploitation. *Molecular informatics* **2010**, *29* (6−7), 476−488.
(38) Cerqueira, N. M.; Gesto, D.; Oliveira, E. F.; Santos-Martins, D.; Brás, N. F.; Sousa, S. F.; Fernandes, P. A.; Ramos, M. J. Receptor-based virtual screening protocol for drug discovery. *Arch. Biochem. Biophys.* **2015**, *582*, 56−67.
(39) Pihan, E.; Kotev, M.; Rabal, O.; Beato, C.; Diaz Gonzalez, C. Fine tuning for success in structure-based virtual screening. *Journal of Computer-Aided Molecular Design* **2021**, *35*, 1195−1206.
(40) Merk, H.; Amran-Gealia, T.; Finkelmeier, D.; Kohl, C.; Pichota, I.; Stern, N.; Rupp, S.; Goldblum, A.; Burger-Kentischer, A. Human-based immune responsive in vitro infection models for validation of novel TLR4 antagonists identified by computational discovery. *Microorganisms* **2022**, *10* (2), 243.
(41) El-Atawneh, S.; Goldblum, A. Iterative stochastic elimination for discovering hits and leads. *Chim. Oggi* **2017**, *35* (5), 41−46.
(42) Deng, Q.; Hu, J.; Wang, L.; Liu, Y.; Guo, Y.; Xu, T.; Pu, X. Probing impact of molecular structure on bulk modulus and impact sensitivity of energetic materials by machine learning methods. *Chemometrics and Intelligent Laboratory Systems* **2021**, *215*, 104331.
(43) Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**, *16* (5), 412−424.
(44) Boughorbel, S.; Jarray, F.; El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PloS one* **2017**, *12* (6), No. e0177678.
(45) Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics* **2020**, *21* (1), 1−13.
(46) Chicco, D.; Tötsch, N.; Jurman, G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining* **2021**, *14* (1), 13.
(47) Tharwat, A. Classification assessment methods. *Applied computing and informatics* **2021**, *17* (1), 168−192.
(48) Bekkar, M.; Djemaa, H. K.; Alitouche, T. A. Evaluation measures for models assessment over imbalanced data sets. *J. Inf. Eng. Appl.* **2013**, *3* (10), 27−38.
(49) Juba, B.; Le, H. S. Precision-recall versus accuracy and the role of large data sets. *Proceedings of the AAAI conference on artificial intelligence* **2019**, *33*, 4039−4048.
(50) Callegari, D.; Pala, D.; Scalvini, L.; Tognolini, M.; Incerti, M.; Rivara, S.; Mor, M.; Lodola, A. Comparative analysis of virtual screening approaches in the search for novel EphA2 receptor antagonists. *Molecules* **2015**, *20* (9), 17132−17151.

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720

(51) Lopes, J. C. D.; Dos Santos, F. M.; Martins-José, A.; Augustyns, K.; De Winter, H. The power metric: a new statistically robust enrichment-type metric for virtual screening applications with early recovery capability. *J. Cheminform.* **2017**, *9* (1), 7.
(52) Fawcett, T. An introduction to ROC analysis. *Pattern recognition letters* **2006**, *27* (8), 861−874.
(53) Brzezinski, D.; Stefanowski, J. Prequential AUC: properties of the area under the ROC curve for data streams with concept drift. *Knowledge and Information Systems* **2017**, *52* (2), 531−562.
(54) Hawkins, P. C.; Warren, G. L.; Skillman, A. G.; Nicholls, A. How to do an evaluation: pitfalls and traps. *Journal of computer-aided molecular design* **2008**, *22* (3−4), 179−190.
(55) Ojala, M.; Garriga, G. C. Permutation tests for studying classifier performance. *J. Machine Learning Res.* **2010**, *11* (6), 1833−1863.
(56) Liu, Y. J.; Kyne, M.; Wang, C.; Yu, X. Y. Data mining in Raman imaging in a cellular biological system. *Computational and Structural Biotechnology Journal* **2020**, *18*, 2920−2930.
(57) Rayan, A.; Falah, M.; Raiyn, J.; Da'adoosh, B.; Kadan, S.; Zaid, H.; Goldblum, A. Indexing molecules for their hERG liability. *Eur. J. Med. Chem.* **2013**, *65*, 304−314.
(58) Keshavarz, M. H.; Seif, F. Improved approach to predict the power of energetic materials. *Propellants, Explosives, Pyrotechnics* **2013**, *38* (5), 709−714.
(59) Explosives Database. http://expdb.chm.uri.edu.
(60) AIST. https://www.aist.go.jp.
(61) Nazari, B.; Keshavarz, M. H.; Hamadanian, M.; Mosavi, S.; Ghaedsharafi, A. R.; Pouretedal, H. R. Reliable prediction of the condensed (solid or liquid) phase enthalpy of formation of organic energetic materials at 298 K through their molecular structures. *Fluid Phase Equilib.* **2016**, *408*, 248−258.
(62) Shivanyuk, A. N.; Ryabukhin, S. V.; Tolmachev, A.; Bogolyubsky, A. V.; Mykytenko, D. M.; Chupryna, A. A.; Kostyuk, A. N. Enamine real database: Making chemical diversity real. *Chim. Oggi* **2007**, *25* (6), 6−7.

https://doi.org/10.1021/acsomega.4c01070
ACS Omega 2024, 9, 42709−42720