Journal of Cheminformatics

RESEARCH Open Access

Integrating QSAR modelling with reinforcement learning for Syk inhibitor discovery

Maria Zavadskaya[1], Anastasia Orlova[1], Andrei Dmitrenko[1*] and Vladimir Vinogradov[1]

Abstract

Spleen tyrosine kinase (Syk) is a crucial mediator of inflammatory processes and a promising therapeutic target for the management of autoimmune disorders, such as immune thrombocytopenia. While several Syk inhibitors are known to date, their efficacy and safety profiles remain suboptimal, necessitating the exploration of novel compounds. The study introduces a novel deep reinforcement learning strategy for drug discovery, specifically designed to identify new Syk inhibitors. The approach integrates quantitative structure–activity relationship (QSAR) predictions with generative modelling, employing a stacking-ensemble model that achieves a correlation coefficient of 0.78. From over 78,000 molecules generated by this methodology, we identified 139 promising candidates with high predicted potency, binding affinity and optimal drug-likeness properties, demonstrating structural novelty while maintaining essential Syk inhibitor characteristics. Our approach establishes a versatile framework for accelerated drug discovery, which is particularly valuable for the development of rare disease therapeutics.

Scientific contribution

The study presents the first application of QSAR-guided reinforcement learning for Syk inhibitor discovery, yielding structurally novel candidates with predicted high potency. The presented methodology can be adapted for other therapeutic targets, potentially accelerating the drug development process.

Keywords Generative design, Drug discovery, QSAR, Machine learning, Syk inhibitors, Reinforcement learning

Introduction

Spleen tyrosine kinase (Syk) is an intracellular nonreceptor protein that belongs to the tyrosine kinase family. Its expression is observed in various immune cells involved in mediating inflammatory responses, including B and T cells, fibroblast-like synoviocytes, and tissue macrophages [1, 2]. Activated Syk triggers a cascade of intracellular signalling pathways, leading to the activation of transcription factors and the expression of genes that perform various biological functions [3, 4].

Syk hyperactivation is observed in many autoimmune, allergic and autoinflammatory diseases, as well as some types of cancer and cardiovascular diseases [3, 5, 6]. This broad involvement in disease pathology makes Syk an attractive therapeutic target for the development of novel drugs [7, 8]. One of the diseases most in need of new Syk inhibitor structures is immune thrombocytopenia (ITP). ITP is a rare autoimmune disorder characterized by a low platelet count in the blood [9]. Although various treatment options exist, Syk inhibitors represent one of the most promising and potentially long-term effective therapeutic approaches [10].

*Correspondence:
Andrei Dmitrenko
dmitrenko@scamt-itmo.ru
[1] Center for AI in Chemistry, ITMO University, Lomonosova St. 9, St. Petersburg 197101, Russia

Currently, several Syk inhibitors are undergoing various stages of clinical trials [6]. Despite their potential, the development of potent Syk inhibitors poses certain challenges ranging from avoiding off-target effects on related kinases to the lack of efficacy seen in clinical trials [8, 11]. Clinical trials have reported instances of hypertension, neutropenia, and diarrhea associated with its use [12]. To date, the only Syk inhibitor approved for the treatment of ITP is fostamatinib [13]. However, its development and clinical application have been hindered by safety concerns and inconsistent efficacy data [14]. These challenges underscore the need for further optimization of Syk inhibitors, including improving selectivity, enhancing drug-like properties, and advancing our understanding of disease mechanisms. In this regard, the search for more effective and safer Syk inhibitors is still underway [3].

In recent years, various computational methods and machine learning approaches have become indispensable tools in biomedical research and pharmaceutical development [15–19]. Advancements in these methods have led scientists to focus on the search for new molecules through techniques such as quantitative structure–activity relationship (QSAR) modelling, pharmacophore modelling and molecular docking [20–23]. These methods significantly accelerate the development of novel compounds for targeted therapy by enabling the prediction of the biological activity of untested molecules [18].

Broadly, there are two main computational approaches for identification of new compounds with potential biological activity: screening existing drug-like databases or data-driven generative design of small molecules with desired properties [24] These approaches are closely interrelated since the models utilized for screening chemical databases are also leveraged for the *in silico* assessment of generated structures for the presence of the desired biological activity [25]. In particular, generative molecular design leverages a variety of machine learning methodologies, including generative adversarial networks (GANs), variational autoencoders (VAEs), and transformer-based approaches [19, 26–28]. Each one offers unique advantages: GANs generate structurally diverse molecules, VAEs enable latent space manipulation, and recurrent neural networks (RNNs) excel at sequence-based molecular generation [29].

A special place in this field is occupied by the reinforcement learning (RL) strategy. This approach is particularly effective for developing targeted inhibitors, as it facilitates the fine-tuning of molecular properties to meet the desired criteria [30]. Recent advancements in RL-based *de novo* molecular generation, such as the fragment-based RL forward synthesis approach proposed by Cai et al. [31], have underscored the potential of RL in drug discovery. This method optimizes molecular candidates by growing them fragment-by-fragment via curated reaction templates, ensuring the synthetic feasibility, drug-likeness, and target affinity of the generated molecules. Notably, this approach successfully identified new drug molecules that were successfully synthesized and exhibited improved activity compared with known compounds.

Available data on recognized Syk inhibitors enables the application of the approaches described above for the development of novel compounds exhibiting inhibitory activity towards Syk [32]. Previously, researchers have conducted several virtual screenings of current drug-like databases via pharmacophore modelling and molecular docking [33–35]. For example, Wang et al. [35] combined computer-based screening with *in vitro* tests to identify substances from the traditional Chinese medicine database. This effort led to the discovery of Tanshinone I, a promising compound that is now undergoing further study [36, 37]. In another study, Xie et al. [38] created a 3D pharmacophore model based on known Syk inhibitors and filtered potential candidates using Lipinski's rule and molecular docking. Among the 30 tested substances, 6 showed good ability to block Syk activity. Additionally, classical machine learning approaches for identifying Syk inhibitors were also considered [39].

Despite these advancements, the generative de novo design of Syk inhibitors has not been previously employed. Inspired by the success of RL-based methodologies, this study presents an approach to the design of new Syk inhibitors by enhancing the FREED + + deep reinforcement learning model (Fig. 1). We used FREED++ as the baseline generative model because of its multiparameter optimization capabilities, specifically its ability to incorporate docking scores – a critical metric for effective inhibitor design. Furthermore, the model's fragment-based molecular construction strategy substantially increases the likelihood of synthetic accessibility, another important factor for later stages of drug development. We adapted the reward function of FREED++ to explicitly consider the target properties of Syk inhibitors, advancing the capabilities of the original generative approach. For that, we trained several machine learning models to predict the negative log of the IC50 and implemented a stacking ensemble to achieve $R^2 = 0.78$ on the test set. Notably, these results establish a new state of the art for the prediction of Syk inhibitor biological activity. Our research not only demonstrates the application of cutting-edge generative models to Syk inhibitor design but also

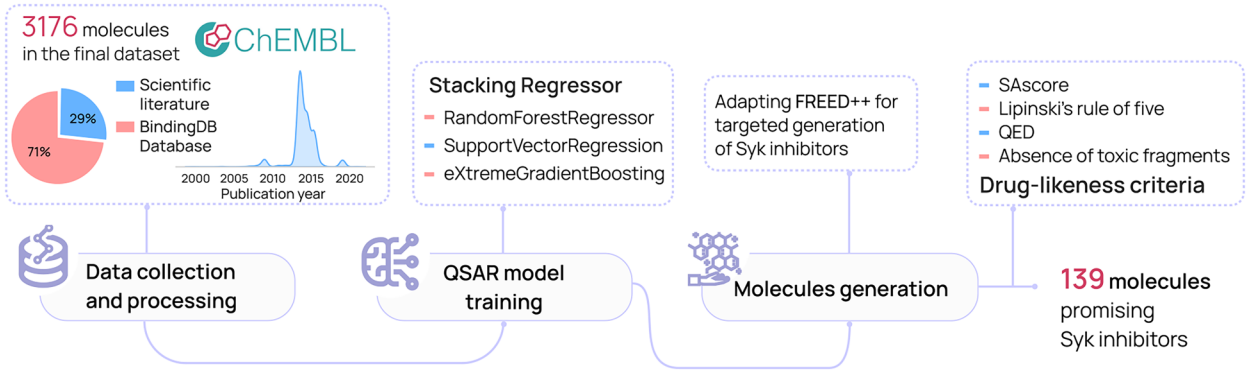

Fig. 1 Computational workflow

introduces a methodology for adapting and improving generative algorithms to address specific therapeutic targets, thereby pushing the boundaries of computational drug discovery.

Materials and methods

Data collection and processing

A dataset of 3,513 Syk inhibitors with an experimentally determined half maximal inhibitory concentration value (IC50) was obtained from the ChEMBL database [40] (target identifier: CHEMBL2599). Within this dataset, 71% of the samples were retrieved from the BindingDB database [41], mainly comprising patent literature, while the remaining 29% were drawn from scientific literature. After preprocessing, including duplicate and outlier removal and inaccurate activity data filtering, a total of 3176 molecules were retained. Duplicated compounds were identified on the basis of SMILES representations. For duplicates with multiple IC50 measurements, the values were averaged if all were within 10% of the median. If the variance exceeded 10% of the median, the lowest IC50 was retained for each compound to avoid the exclusion of potentially potent inhibitors. For machine learning model development, IC50 values were converted to pIC50 values by applying the negative logarithm, ensuring a normalized data distribution suitable for predictive modelling. The distribution of pIC50 values in the final dataset is provided in Sect. 1 of the Supplementary Materials.

This curated dataset comprises unique molecules, of which 1,642 are highly potent inhibitors (IC50 < 50 nM), 999 are moderately active (50 nM < IC50 < 500 nM), and 535 are lowly active (IC50 > 500 nM). To represent molecular structures, various methods, both novel and generally accepted, were evaluated via the Pycaret autoML framework to identify the optimal approach. The obtained structure–activity data formed the foundation for developing a machine learning regression model to predict pIC50 values.

Bioactivity prediction

The following classical machine learning models have been employed to predict the inhibitory ability of drug molecules: Random Forest Regression (RFR), Hist Gradient Boosting (HGB), eXtreme Gradient Boosting (XGB) and Support Vector Regression (SVR). These models were selected due to their proven efficacy in addressing regression tasks within the pharmaceutical industry [42, 43]. Parameter optimization for these models was conducted utilizing the Optuna framework [44]. A comprehensive description of this process is provided in Sect. 3 of the Supplementary Materials.

RFR, HGB and XGB refer to the ensemble learning methods. These methods involve constructing multiple decision trees and combining them to achieve improved predictive performance compared with individual decision trees. SVR, a regression analysis method based on support vectors, is adept at handling non-linear dependencies and high-dimensional data [45], making it well-suited for accurately predicting efficacy based on long binary vectors – molecular fingerprints.

To further improve the predictive performance, we implemented a stacking ensemble approach. This method combines multiple base models into a meta-regressor leveraging the individual strengths of each model to create a more robust and accurate predictive framework. The top-performing algorithms from our initial evaluation were selected as the base models, while a standard linear regression model was employed as the final estimator in the ensemble. By aggregating the predictions of

diverse models, the stacking approach enhances overall accuracy and robustness.

To evaluate and compare the predictive performance of all models, individual models (RFR, HGB, XGB, SVR) as well as a stacking ensemble with fivefold cross-validation (CV) were applied to ensure robust performance estimation. The performance metrics used for evaluation were the coefficient of determination (R-squared) and the mean squared error (MSE).

Molecules generation

We adopted an RL approach for de novo drug molecule generation, prioritizing desired chemical space and multi-property optimization [46, 47]. This was achieved by incorporating docking score, drug-likeness, and bioactivity (specifically pIC50) into the RL reward function.

3fqs protein from the PDB database [48] was selected as the target for generation of new inhibitor molecules. This protein is known to be complexed with R406, an active metabolite that previously served as the basis for the most commonly used agent for inhibiting Syk, Fostamatinib [13].

Potential ligand binding sites within the protein structure were identified using the following procedure: first, the three-dimensional structure of the ligand was extracted from the corresponding PDB file; then, the centroid of the bounding parallelepiped was calculated by averaging the coordinates of all ligand atoms. The dimensions of the parallelepiped along each coordinate axis were estimated by adding the maximum deviation of the ligand atom coordinates from the respective centroid coordinate plus 4 Å. A similar method for preparing potential ligand binding sites was used by the authors of FREED++ [49].

The CReM-ZINC fragment library from FREED++ was used as the source library of molecular fragments. The reward functions used for structure generation employed the following parameters: lipophilicity in the form of octanol–water partition coefficient (logP), the number of heavy atoms in the molecule (HeavyAtomCount), the number of hydrogen bond acceptors (NumHAcceptors) and donors (NumHDonors), as well as filters to exclude potentially toxic and undesirable pharmacophoric groups and fragments (PAINS, SureChEMBL, Glaxo). The number of epochs was set to 200, which consistently resulted in around 13 thousand generated molecules.

To ensure the generation of potent Syk inhibitor structures, we upgraded the reward function in a way that it takes into account compound bioactivity. This adaptation, which uses the pretrained pIC50 prediction model, was incorporated into the FREED++ reward function as an additional objective parameter.

Evaluation metrics for the generated molecules

After generation, a comprehensive assessment and filtering process was applied. First, any structures that caused errors in the RDKit software were eliminated to ensure data integrity. For the remaining molecules, several key properties were calculated to evaluate their potential as drug candidates: synthetic accessibility score [50] (SAscore < 6), quantitative estimate of drug-likeness [51] (QED > 0.67), and absence of toxic fragments (the number of toxic groups equals 0). SAscore estimates the ease of molecule synthesis, with lower scores indicating simpler synthetic routes. QED score reflects how closely a compound's properties match those of known drugs, with higher scores indicating greater drug-likeness. Finally, the absence of toxic fragments ensures the exclusion of molecules containing substructures associated with known toxicity, thus enhancing the safety profile of potential drug candidates. The thresholds for these selection criteria were established based on the parameters utilized in the ADMETLAB 3.0 platform [52].

To evaluate the effectiveness of the generated molecules, the developed predictive model for Syk inhibition efficiency (pIC50) was applied. A pIC50 threshold of 7.40 was chosen to prioritize molecules with greater inhibitory potency than existing marketed drugs [53]. Furthermore, the docking score (DS), calculated during the generation process, was used as an additional filtering criterion, with only molecules achieving DS values below -7 kcal/mol being retained. This threshold aligns with established literature, where effective Syk inhibitors consistently show DS values between -7 kcal/mol and -10 kcal/mol [54–57], and represents a standard cutoff for identifying compounds with promising binding affinity.

Results and discussion

Predictive QSAR model

We developed a machine learning regression model to predict the inhibitory potency from molecular structure. This approach, known as QSAR modelling, utilizes the half-maximal inhibitory concentration value (pIC50) as a quantitative measure of inhibitory potency. The dataset was split into training and test sets in a 4:1 ratio, ensuring similar distributions of molecular structures and pIC50 values, as detailed in Sect. 2 of the Supplementary Materials.

To construct the QSAR model, we evaluated five molecular representation methods using the PyCaret autoML framework. Among these, extended-connectivity fingerprints (ECFPs) demonstrated the best performance metrics (Table 1). Molecular structures were encoded as ECFPs using the RDKit cheminformatics software package. This encoded structure–activity data formed the

Table 1 Comparison of molecules representation methods using PyCaret autoML framework

Method	Fingerprint category	LightGBM		Random Forest		SVR	
		Q^2	MSE CV	Q^2	MSE CV	Q^2	MSE CV
ECFP [58]	Circular	**0.71**	**0.37**	**0.70**	**0.38**	**0.72**	**0.34**
Mol2vec [59]	Substructure	0.64	0.46	0.60	0.51	0.47	0.66
MACCS [60]		0.63	0.47	0.61	0.50	0.60	0.49
PubChem fingerprints [61]		0.66	0.44	0.65	0.45	0.62	0.47
MAP4 [62]	String	0.64	0.46	0.61	0.50	0.69	0.38

The bold values highlight the best results within each column, corresponding to the specific model and evaluation metric

Table 2 Performance metrics of the base models on the test and train dataset with optimized hyperparameters

Models	Train dataset		Test dataset	
	MSE CV	Q^2	MSE	R^2
HistGradientBoosting	0.38 ± 0.04	0.70 ± 0.03	0.37	0.71
RandomForestRegressor	0.38 ± 0.04	0.69 ± 0.03	0.35	0.73
SupportVectorRegression	0.34 ± 0.03	0.72 ± 0.03	0.36	0.72
eXtremeGradientBoosting	**0.32 ± 0.05**	0.74 ± 0.04	0.32	0.75
StackingRegressor	0.33 ± 0.02	**0.74 ± 0.02**	**0.27**	**0.78**

The bold values highlight the best results within each column, corresponding to the specific dataset and evaluation metric

foundation for QSAR modelling, where pIC50 values are used as the target variable to quantify inhibitory potency.

Four most robust and widely used ML models were chosen for constructing the predictive model, namely, HGB, RF, XGB and SVR. Initially, all models were trained with their default parameters. Subsequently, hyperparameter optimization using the Optuna library significantly improved the performance of all four algorithms. The performance metrics of the optimized models are outlined in Table 2.

Based on the optimization results, the top-performing models—RF, XGB, and SVR—were selected as base learners for the stacking ensemble. We implemented a stacking ensemble technique by combining these base models. A meta-regressor, specifically a standard linear regression model, was used as the final estimator in the stacking ensemble (Fig. 2). The stacking regressor was implemented using the StackingRegressor class from scikit-learn, with fivefold CV applied during training to ensure robust performance estimation. After CV, the stacking regressor was trained on the entire training set, achieving the best predictive performance on the test set: the lowest MSE equal of 0.27, and the highest R-squared of 0.78. Based on these results, we opted for StackingRegressor as our primary model for estimating pIC50 values of novel Syk inhibitors.

Generation of new Syk inhibitors

Experiments on the generation of new Syk inhibitors involved three distinct approaches: (1) baseline

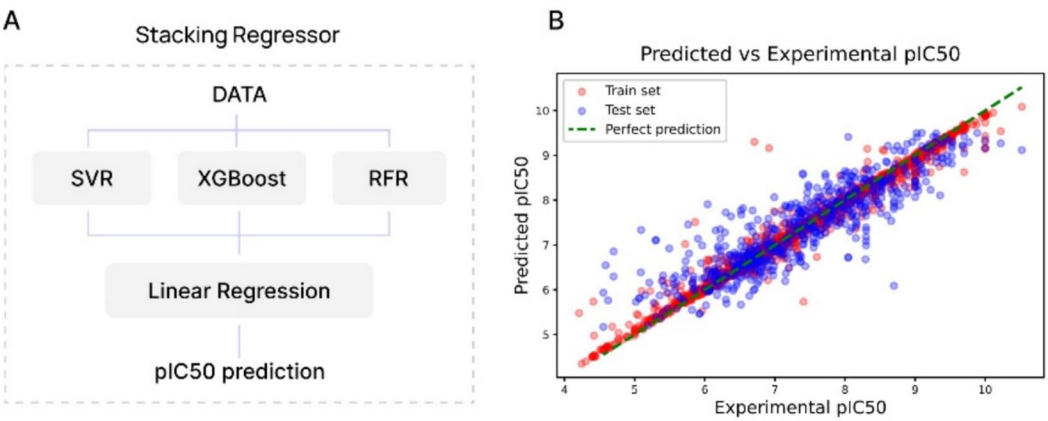

Fig. 2 Development of machine learning models. **A** Stacking regressor architecture. **B** Scatter plot of the stacking regressor on the train and test sets

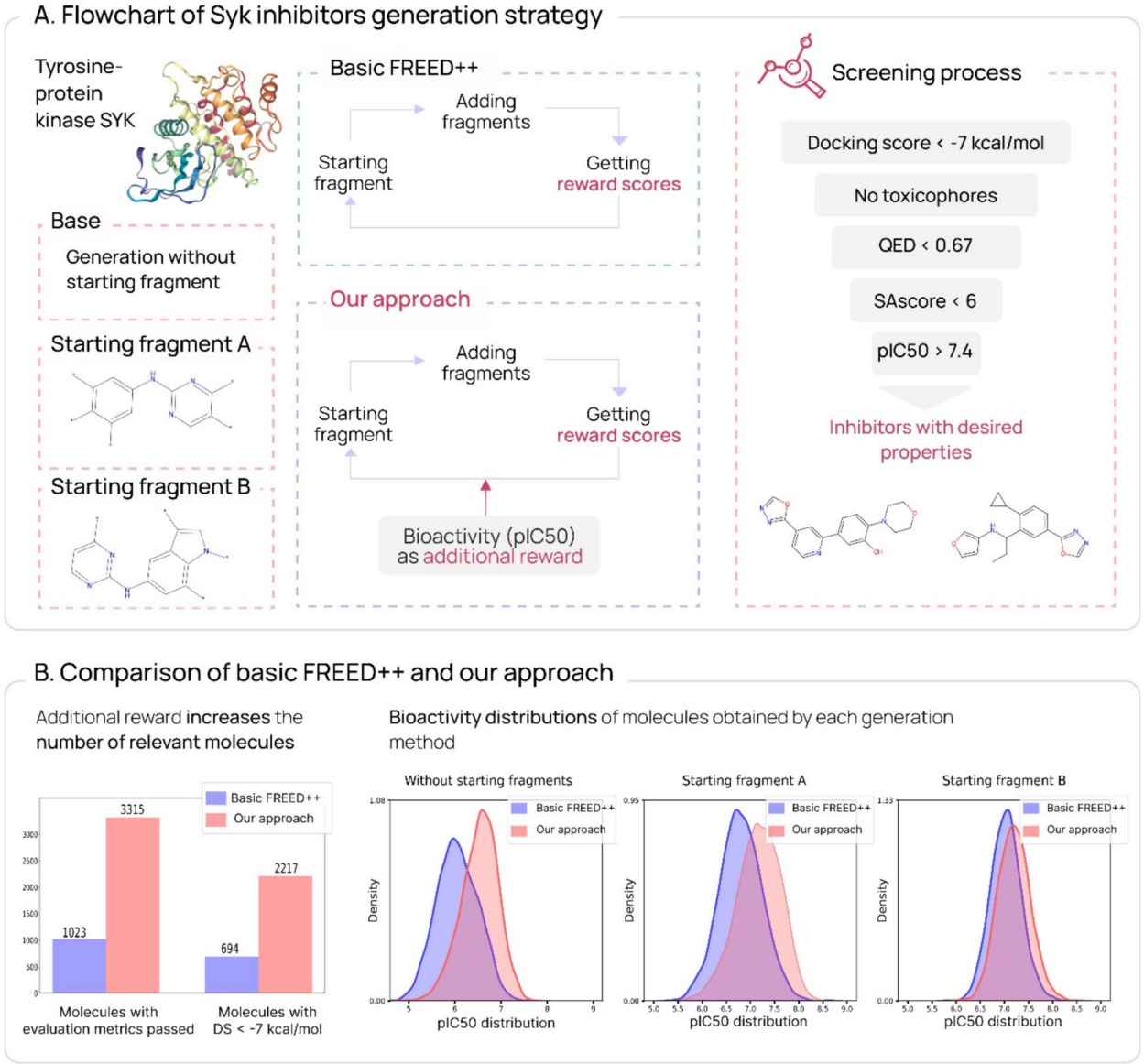

Fig. 3 **A** Flowchart of our inhibitors generation strategy. Starting fragments used for molecule generation. The strategy for selecting fragments is described in more detail in Sect. 4 of the Supplementary materials. **B** Comparison of basic and improved FREED++

FREED++ model with standard reward function parameters, (2) our approach incorporating the pre-trained pIC50 prediction model into the FREED++ reward function, and (3) the utilization of starting fragments regardless of the aforementioned generative models (Fig. 3A). The selection of starting fragments was based on a dataset of known inhibitors, as described in Sect. 4 of the Supplementary materials.

We compared the outcomes of the generation experiments by assessing the number of molecules passing the screening criteria (Sect. "Evaluation metrics for the generated molecules") after eliminating invalid and duplicate structures. The outcomes of this comparative analysis are summarized in Table 3. Incorporating the pre-trained pIC50 prediction model into the reward function not only preserved molecular uniqueness but also increased the number of valid molecules generated. Our approach demonstrated better performance across all evaluation criteria compared with the baseline version, yielding a greater number of molecules with drug-like properties and improved pIC50 and DS values. Notably, the total

Table 3 The generation performance of basic FREED++ and our approach

	FREED++	Our approach
Valid molecules	17,689	**20,313**
DockingScore < -7	12,410	**13,821**
Drug-like properties passed	2249	**3774**
Active molecules (pIC50 > 7.4 and DS < -7)	629	**2684**
Active ratio (%)	3.56	**13.21**

The bold values highlight the best results within each column, corresponding to the specific approach and comparison criterion

Table 4 The number of generated molecules passing screening thresholds

	Valid drug-like molecules with pIC50 > 7.4 and DS < − 7 kcal/mol				
	Base	**Start. Fr. A**	**Start. Fr. B**	**Total**	**Tanimoto similarity**
FREED++	3	65	1	69	0.35 ± 0.12
Our approach	**17**	43	**12**	**72**	**0.27** ± 0.11

The bold values highlight the best results within each column, corresponding to the specific approach and evaluation metric

percentage of recruited active molecules increased from 3.56% in the baseline model to 13.21% with our approach.

Further analysis focused on the subset of valid, drug-like molecules with pIC50 > 7.40 and DS < − 7 kcal/mol, which represent the most promising Syk inhibitors. The results from all six experiments are presented in Table 4. While the baseline FREED++ model generated a larger quantity of molecules with scaffold A, our approach increased the total number of promising inhibitors and significantly enhanced their structural diversity. This improvement was evidenced by the number of molecules of different scaffolds and the Tanimoto similarity between the molecular fingerprints of the generated molecules (0.274 by our approach vs. 0.354 by FREED++).

Comparing our approach to the baseline model, we also observed a clear shift in the pIC50 distribution toward higher values, regardless of the starting fragment (Fig. 3B). This shift highlights the improved potency of the generated molecules. The statistical significance of these results was confirmed by the Mann–Whitney tests (see Sect. 5 of the Supplementary materials).

The integration of the QSAR model in the reward function has proven effective for the de novo design of potential Syk inhibitors. In the future, this approach can be further refined: as more experimental data on selectivity and off-target effects become available, the reward function can be modified to include those additional desirable properties of Syk inhibitors. More generally, the flexibility of our approach allows for fine-tuning the generative process to produce molecules of specific therapeutic requirements.

Importantly, the proposed methodology is not limited to Syk inhibitors. In principle, it can be generalized to a wide range of therapeutic targets provided the biological activity data is available for training. However, it is important to recognize the fundamental limitations of this methodology. The quality of the generated molecules heavily depends on the accuracy of the underlying QSAR model and the ultimate structure of the reward function. For instance, the QSAR model's performance is constrained by the composition and biases of the training data, which comprised patent-derived and literature-derived data. Both sources primarily report successful molecules, as negative results are rarely published, leading to a model trained mostly on "successful" candidates. This bias may hinder generalization to less promising molecules, potentially causing overly optimistic predictions. Furthermore, insights from the scaffold-clustering analysis conducted in this study (see Sect. 4 of the Supplementary Materials) revealed that Syk inhibitors in the training data tend to share common structural fragments. This structural homogeneity poses a challenge for the QSAR model, as it may struggle to make reliable predictions for molecules with significantly different scaffolds. Generative process may prioritize molecules with familiar scaffolds, potentially reducing the novelty and diversity of the generated candidates. Finally, predictions made by the proposed methodology require comprehensive experimental validation, as no computational approach is capable of capturing all nuances of drug-target interactions.

To address these limitations, future work could focus on mitigating biases in the training data. For instance, providing models with essential negative examples could enhance their ability to predict properties across a broader range of molecular outcomes. Additionally, the reward function could be refined to encourage generation of molecules with novel scaffolds. Most importantly, we intend to conduct in vitro characterization of the generated compounds to assess their potential for therapeutic applications.

Property analysis of potential inhibitors

Through a rigorous screening process of the 78,012 generated molecules, we successfully identified 139 compounds (see Sect. 6 of the Supplementary Materials) that satisfied the predefined selection criteria outlined in Sect. "Evaluation metrics for the generated molecules".

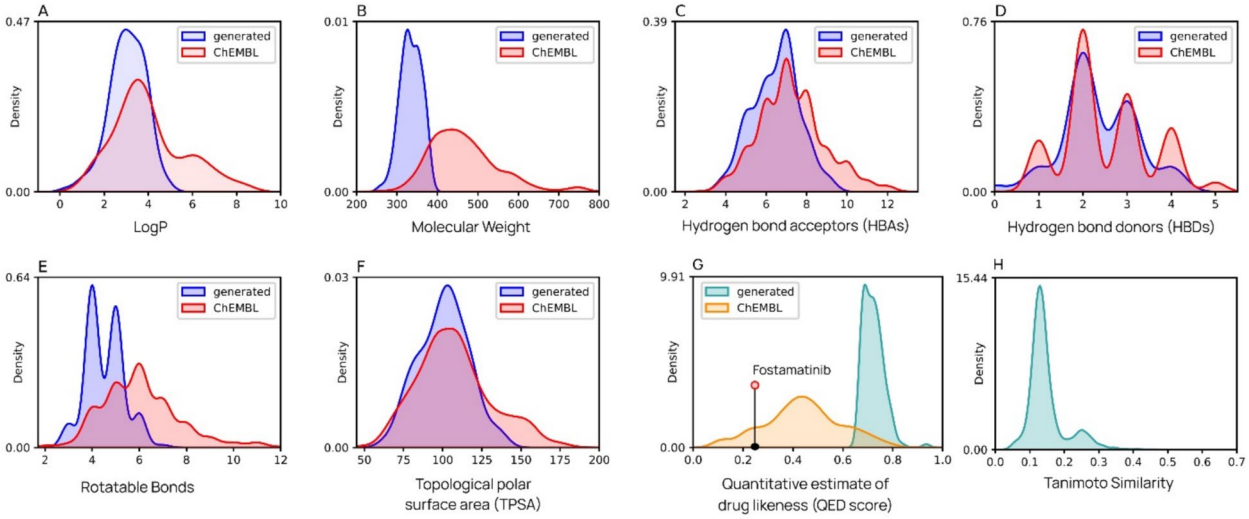

Fig. 4 Distribution of the calculated molecular properties of generated molecules (blue) and ChEMBL molecules (red): **A** Log P, **B** Molecular weight, **C** number of HBAs, **D** number of HBDs, **E** number of rotatable bonds, **F** TPSA, **G** QED, **H** Tanimoto similarity distribution between the generated set and known compounds

To further validate the molecular properties of the generated candidates, we compared them to known inhibitors with comparable potency (pIC50 > 7.4) from the dataset used to train the QSAR model. Key investigated parameters included the partition coefficient (LogP), molecular weight (MW), number of hydrogen bond acceptors and donors (HBA and HBD), topological polar surface area (TPSA) and number of rotatable bonds (RotatableBonds). Distributions of most molecular descriptors for the generated molecules aligned well with those of experimentally confirmed compounds, except for the molecular weight and the number of rotatable bonds, which both shifted towards lower values (Fig. 4A–F). However, this reduction may confer advantageous pharmacokinetic properties and enhanced bioavailability for potential drug candidates [63, 64].

To assess the structural novelty of the generated candidate molecules in comparison with previously reported compounds, a set of 122 molecules was selected based on the criteria employed for candidate screening. The analysis of Tanimoto similarity between the two sets revealed that 98% of the molecular pairs showed similarity coefficients lower than 0.3 (Fig. 4H). These findings highlight that the generated molecules exhibit structural novelty while maintaining drug-like properties comparable to those of known compounds.

One of the main concerns with Syk inhibitors undergoing clinical trials is the risk of adverse events. To explore this issue, we assessed fostamatinib, a clinically tested Syk inhibitor, using the same criteria employed in the screening of the generated molecules. Our analysis revealed that fostamatinib has a low quantitative estimation of drug-likeness (QED) score of 0.256. In contrast, our set of generated molecules demonstrates favourable characteristics in these aspects (Fig. 4G).

Conclusion

In this study, we successfully applied a novel approach combining deep reinforcement learning with QSAR predictive model to design new potential Syk inhibitors. This method enabled us to generate a set of 139 promising candidate molecules with high predicted potency and favourable drug-like properties. Importantly, the generated compounds exhibited structural novelty while maintaining molecular characteristics similar to known Syk inhibitors, potentially addressing the limitations of existing drugs such as fostamatinib.

By incorporating QSAR predictions into the generative process, we effectively guided molecular generation toward the desired chemical space, resulting in an increased yield of high-quality drug candidates. This methodology not only accelerates the early stages of drug discovery for Syk inhibitors but also presents a versatile framework that can be adapted for other therapeutic targets. Its applicability is particularly promising for orphan diseases, such as immune thrombocytopenia, where novel therapies are urgently needed. Our future work will focus on experimental validation of the top candidate molecules and further refinements of the generative model. Additionally, we look forward to extending our approach to other therapeutic targets, such as Bruton's tyrosine kinase or Forkhead box M1.

Abbreviations

Syk	Spleen tyrosine kinase
QSAR	Quantitative structure–activity relationship
ITP	Immune thrombocytopenia
GANs	Generative adversarial networks
VAEs	Variational autoencoders
RNNs	Recurrent neural networks
RL	Reinforcement learning
IC50	Half maximal inhibitory concentration value
RFR	Random Forest Regression
HGB	Hist Gradient Boosting
XGB	EXtreme Gradient Boosting
SVR	Support Vector Regression
CV	Cross-validation
MSE	Mean squared error
R-squared	Coefficient of determination
SAscore	Synthetic accessibility score
QED	Quantitative estimate of drug-likeness
DS	Docking score
ECFPs	Extended-connectivity fingerprints

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13321-025-00998-2.

Additional file 1. Distribution of the pIC50 values of Syk inhibitors in the final dataset. Hyperparameter Optimization. Scaffold-clustering analysis. Statistical tests for comparing approaches to molecule generation. Generated molecules.

Acknowledgements
This study was supported by the Priority 2030 Federal Academic Leadership Program.

Author contributions
Maria Zavadskaya: Writing – original draft, Visualization, Software, Resources, Methodology, Investigation, Formal analysis, Data curation. Anastasia Orlova: Writing – review & editing, Conceptualization, Methodology, Supervision. Andrei Dmitrenko: Writing – review & editing, Validation, Supervision, Formal analysis. Vladimir Vinogradov: Writing – review & editing, Conceptualization, Supervision.

Availability of data and materials
The data used in this study and the codes can be found at https://github.com/ai-chem/Syk_inhibitors.

Declarations

Competing interests
The authors declare no competing interests.

Received: 17 January 2025 Accepted: 26 March 2025
Published online: 15 April 2025

References

1. Zhou Y, Zhang Y, Yu W et al (2023) Immunomodulatory role of spleen tyrosine kinase in chronic inflammatory and autoimmune diseases. Immun Inflamm Dis 11:e934. https://doi.org/10.1002/iid3.934
2. Pamuk ON, Tsokos GC (2010) Spleen tyrosine kinase inhibition in the treatment of autoimmune, allergic and autoinflammatory diseases. Arthritis Res Ther 12:222. https://doi.org/10.1186/ar3198
3. Wang Z, Qu S, Yuan J et al (2023) Review and prospects of targeted therapies for spleen tyrosine kinase (SYK). Bioorg Med Chem 96:117514. https://doi.org/10.1016/j.bmc.2023.117514
4. Mócsai A, Ruland J, Tybulewicz VLJ (2010) The SYK tyrosine kinase: a crucial player in diverse biological functions. Nat Rev Immunol 10:387–402. https://doi.org/10.1038/nri2765
5. Li M, Wang P, Zou Y et al (2023) Spleen tyrosine kinase (SYK) signals are implicated in cardio-cerebrovascular diseases. Heliyon 9:e15625. https://doi.org/10.1016/j.heliyon.2023.e15625
6. Patton JT, Woyach JA (2024) Targeting the B cell receptor signaling pathway in chronic lymphocytic leukemia. Semin Hematol. https://doi.org/10.1053/j.seminhematol.2024.04.002
7. Dummer W, Markovtsov VV, Tong S et al (2020) Clinical trial to evaluate an approved ITP therapy targeting spleen tyrosine kinase (SYK) for prevention and treatment of COVID-19 related complications. Blood 136:35. https://doi.org/10.1182/blood-2020-141045
8. Liu D, Mamorska-Dyga A (2017) Syk inhibitors in clinical development for hematological malignancies. J Hematol Oncol. https://doi.org/10.1186/s13045-017-0512-1
9. Al-Samkari H, Neufeld EJ (2023) Novel therapeutics and future directions for refractory immune thrombocytopenia. Br J Haematol 203:65–78. https://doi.org/10.1111/bjh.19078
10. Liu XG, Hou Y, Hou M (2023) How we treat primary immune thrombocytopenia in adults. J Hematol Oncol 16:4. https://doi.org/10.1186/s13045-023-01401-z
11. Loos NHC, Sparidans RW, Heydari P et al (2024) The ABCB1 and ABCG2 efflux transporters limit brain disposition of the SYK inhibitors entospletinib and lanraplenib. Toxicol Appl Pharmacol 485:116911. https://doi.org/10.1016/j.taap.2024.116911
12. Genovese MC, Kavanaugh A, Weinblatt ME et al (2011) An oral Syk kinase inhibitor in the treatment of rheumatoid arthritis: a three-month randomized, placebo-controlled, phase II study in patients with active rheumatoid arthritis that did not respond to biologic agents. Arthritis Rheum 63:337–345. https://doi.org/10.1002/art.30114
13. Paik J (2021) Fostamatinib: a review in chronic immune thrombocytopenia. Drugs 81:935–943. https://doi.org/10.1007/s40265-021-01524-y
14. Britto J, Holbrook A, Sun H et al (2024) Thrombopoietin receptor agonists and other second-line therapies for immune thrombocytopenia: a narrative review with a focus on drug access in Canada. Clin Invest Med 47:13–22. https://doi.org/10.3138/cim-2024-2569
15. Pushkaran AC, Arabi AA (2024) From understanding diseases to drug design: can artificial intelligence bridge the gap? Artif Intell Rev 57:1–39. https://doi.org/10.1007/s10462-024-10714-5
16. Mouchlis VD, Afantitis A, Serra A et al (2021) Advances in de novo drug design: from conventional to machine learning methods. Int J Mol Sci 22:1676. https://doi.org/10.3390/ijms22041676
17. Obaido G, Mienye ID, Egbelowo OF et al (2024) Supervised machine learning in drug discovery and development: algorithms, applications, challenges, and prospects. Mach Learn Appl 17:100576. https://doi.org/10.1016/j.mlwa.2024.100576
18. Parvatikar PP, Patil S, Khaparkhuntikar K et al (2023) Artificial intelligence: machine learning approach for screening large database and drug discovery. Antiviral Res 220:105740. https://doi.org/10.1016/j.antiviral.2023.105740
19. Gangwal A, Lavecchia A (2024) Unleashing the power of generative AI in drug discovery. Drug Discov Today 29:103992. https://doi.org/10.1016/j.drudis.2024.103992
20. Tropsha A, Isayev O, Varnek A et al (2023) Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov 23:141–155. https://doi.org/10.1038/s41573-023-00832-0
21. Matada GSP, Dhiwar PS, Abbas N et al (2022) Pharmacophore modeling, virtual screening, molecular docking and dynamics studies for the discovery of HER2-tyrosine kinase inhibitors: an in-silico approach. J Mol Struct 1257:132531. https://doi.org/10.1016/j.molstruc.2022.132531
22. Schaller D, Šribar D, Noonan T et al (2020) Next generation 3D pharmacophore modeling. Wiley Interdiscip Rev Comput Mol Sci 10:e1468. https://doi.org/10.1002/wcms.1468
23. Kawsar SMA, Munia NS, Saha S et al (2023) In silico pharmacokinetics, molecular docking and molecular dynamics simulation studies of nucleoside analogs for drug discovery: a mini review. Mini Rev Med Chem 24:1070–1088. https://doi.org/10.2174/0113895575258033231024073521

24. Sadybekov AV, Katritch V (2023) Computational approaches streamlining drug discovery. Nature 616:673–685. https://doi.org/10.1038/s41586-023-05905-z
25. Janet JP, Mervin L, Engkvist O et al (2023) Artificial intelligence in molecular de novo design: integration with experiment. Curr Opin Struct Biol 80:102575. https://doi.org/10.1016/j.sbi.2023.102575
26. Pang C, Qiao J, Zeng X et al (2024) Deep generative models in de novo drug molecule generation. J Chem Inf Model 64:2174–2194. https://doi.org/10.1021/acs.jcim.3c01496
27. Mathivanan JS, Dhayabaran VV, David MR et al (2024) Application of deep learning neural networks in computer-aided drug discovery: a review. Curr Bioinform 19:851–858. https://doi.org/10.2174/0115748936276510231123121404
28. Gangwal A, Ansari A, Ahmad I et al (2024) Generative artificial intelligence in drug discovery: basic framework, recent advances, challenges, and opportunities. Front Pharmacol 15:1331062. https://doi.org/10.3389/fphar.2024.1331062
29. Kanakala GC, Devata S, Chatterjee P et al (2024) Generative artificial intelligence for small molecule drug design. Curr Opin Biotechnol 89:103175. https://doi.org/10.1016/j.copbio.2024.103175
30. Sridharan B, Sinha A, Bardhan J et al (2024) Deep reinforcement learning in chemistry: a review. J Comput Chem 45:1886–1898. https://doi.org/10.1002/jcc.27354
31. Cai H, Chen W, Jiang J et al (2024) Artificial intelligence-assisted optimization of antipigmentation tyrosinase inhibitors: de novo molecular generation based on a low activity lead compound. J Med Chem 67:7260–7275. https://doi.org/10.1021/acs.jmedchem.4c00091
32. Kumar V, Parate S, Danishuddin S et al (2022) 3D-QSAR-based pharmacophore modeling, virtual screening, and molecular dynamics simulations for the identification of spleen tyrosine kinase inhibitors. Front Cell Infect Microbiol 12:909111. https://doi.org/10.3389/fcimb.2022.909111
33. Samanta S, Sk MF, Koirala S et al (2023) Exploring molecular interactions of potential inhibitors against the spleen tyrosine kinase implicated in autoimmune disorders via virtual screening and molecular dynamics simulations. SAR QSAR Environ Res 34:869–897. https://doi.org/10.1080/1062936x.2023.2266364
34. Ali T, Anjum F, Choudhury A et al (2024) Identification of natural product-based effective inhibitors of spleen tyrosine kinase (SYK) through virtual screening and molecular dynamics simulation approaches. J Biomol Struct Dyn. https://doi.org/10.1080/07391102.2023.2218938
35. Wang X, Guo J, Ning Z et al (2018) Discovery of a natural Syk inhibitor from Chinese medicine through a docking-based virtual screening and biological assay study. Molecules 23:3114. https://doi.org/10.3390/molecules23123114
36. Jieensinue S, Zhu H, Li G et al (2018) Tanshinone IIA reduces SW837 colorectal cancer cell viability via the promotion of mitochondrial fission by activating JNK-Mff signaling pathways. BMC Cell Biol 19:21. https://doi.org/10.1186/s12860-018-0174-z
37. Tung MC, Tsai KC, Fung KM et al (2021) Characterizing the structure-activity relationships of natural products, tanshinones, reveals their mode of action in inhibiting spleen tyrosine kinase. RSC Adv 11:2453–2461. https://doi.org/10.1039/d0ra08769f
38. Xie HZ, Li LL, Ren JX et al (2009) Pharmacophore modeling study based on known spleen tyrosine kinase inhibitors together with virtual screening for identifying novel inhibitors. Bioorg Med Chem Lett 19:1944–1949. https://doi.org/10.1016/j.bmcl.2009.02.049
39. Li BK, Cong Y, Yang XG et al (2013) In silico prediction of spleen tyrosine kinase inhibitors using machine learning approaches and an optimized molecular descriptor subset generated by recursive feature elimination method. Comput Biol Med 43:395–404. https://doi.org/10.1016/j.compbiomed.2013.01.015
40. Zdrazil B, Felix E, Hunter F et al (2024) The ChEMBL database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. Nucleic Acids Res 52:D1180–D1192. https://doi.org/10.1093/nar/gkad1004
41. Gilson MK, Liu T, Baitaluk M et al (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053. https://doi.org/10.1093/nar/gkv1072
42. Campos MSR, López DAG, Rivera JAC et al (2022) Bioactivity predictors for the inhibition of *Staphylococcus aureus* quinolone resistance protein. Commun Comput Inf Sci 1685:31–40. https://doi.org/10.1007/978-3-031-20611-5_3
43. Singh K, Ghosh I, Jayaprakash V et al (2024) Building a ML-based QSAR model for predicting the bioactivity of therapeutically active drug class with imidazole scaffold. Eur J Med Chem Rep 11:100148. https://doi.org/10.1016/j.ejmcr.2024.100148
44. Akiba T, Sano S, Yanase T et al (2019) Optuna: a next-generation hyperparameter optimization framework. In: Akiba T (ed) Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining. Association for computing machinery, New York. https://doi.org/10.1145/3292500.3330701
45. Brereton RG, Lloyd GR (2010) Support vector machines for classification and regression. Analyst 135:230–267. https://doi.org/10.1039/b918972f
46. Blaschke T, Engkvist O, Bajorath J et al (2020) Memory-assisted reinforcement learning for diverse molecular de novo design. J Cheminform 12:68. https://doi.org/10.1186/s13321-020-00473-0
47. Korshunova M, Huang N, Capuzzi S et al (2022) Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. Commun Chem 5:129. https://doi.org/10.1038/s42004-022-00733-0
48. RCSB PDB-3FQS: crystal structure of spleen tyrosine kinase complexed with R406. https://www.rcsb.org/structure/3FQS. Accessed 1 Mar 2025.
49. Telepov A, Tsypin A, Khrabrov K et al (2024) FREED++: improving RL agents for fragment-based molecule generation by thorough reproduction. arXiv preprint. https://doi.org/10.48550/arXiv.2401.09840
50. Ertl P, Schuffenhauer A (2009) Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. J Cheminform 1:8. https://doi.org/10.1186/1758-2946-1-8
51. Bickerton GR, Paolini GV, Besnard J et al (2012) Quantifying the chemical beauty of drugs. Nat Chem 4:90–98. https://doi.org/10.1038/nchem.1243
52. Fu L, Shi S, Yi J et al (2024) ADMETlab 3.0: an updated comprehensive online ADMET prediction platform enhanced with broader coverage, improved performance, API functionality and decision support. Nucleic Acids Res 52:W422–W431. https://doi.org/10.1093/nar/gkae236
53. Rolf MG, Curwen JO, Veldman-Jones M et al (2015) In vitro pharmacological profiling of R406 identifies molecular targets underlying the clinical effects of fostamatinib. Pharmacol Res Perspect 3:e00175. https://doi.org/10.1002/prp2.175
54. Hu X, Hu C, Liao L et al (2024) Isoliquiritigenin limits inflammasome activation of macrophage via docking into Syk to alleviate murine non-alcoholic fatty liver disease. Scand J Immunol 100:e13371. https://doi.org/10.1111/sji.13371
55. Marchetti G, Dessì A, Dallocchio R et al (2020) Syk inhibitors: new computational insights into their intraerythrocytic action in *Plasmodium falciparum* malaria. Int J Mol Sci 21:7009. https://doi.org/10.3390/ijms21197009
56. Mansouri M, ElHaddoumi G, Kandoussi I et al (2024) Syk protein inhibitors treatment for the allergic symptoms associated with hyper immunoglobulin E syndromes: a focused on a computational approach. Int J Immunopathol Pharmacol 38:3946320241282030. https://doi.org/10.1177/03946320241282030
57. Wang L, Fang Y, Ma Y et al (2024) A novel natural Syk inhibitor suppresses IgE-mediated mast cell activation and passive cutaneous anaphylaxis. Bioorg Chem 146:107320. https://doi.org/10.1016/j.bioorg.2024.107320
58. Rogers D, Hahn M (2010) Extended-connectivity fingerprints. J Chem Inf Model 50:742–754. https://doi.org/10.1021/ci100050t
59. Jaeger S, Fulle S, Turk S (2018) Mol2vec: unsupervised machine learning approach with chemical intuition. J Chem Inf Model 58:27–35. https://doi.org/10.1021/acs.jcim.7b00616
60. Durant JL, Leland BA, Henry DR et al (2002) Reoptimization of MDL keys for use in drug discovery. J Chem Inf Comput Sci 42:1273–1280. https://doi.org/10.1021/ci010132r
61. Kim S, Chen J, Cheng T et al (2025) PubChem 2025 update. Nucleic Acids Res 53:D1516–D1525. https://doi.org/10.1093/nar/gkae1059
62. Capecchi A, Probst D, Reymond JL (2020) One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. J Cheminform 12:43. https://doi.org/10.1186/s13321-020-00445-4
63. Meanwell NA (2011) Improving drug candidates by design: a focus on physicochemical properties as a means of improving compound disposition and safety. Chem Res Toxicol 24:1420–1456. https://doi.org/10.1021/tx200211v

64. Tian S, Li Y, Wang J et al (2011) ADME evaluation in drug discovery. 9. Prediction of oral bioavailability in humans based on molecular properties and structural fingerprints. Mol Pharm 8:841–851. https://doi.org/10.1021/mp100444g

Publisher's Note