



Computational design of MMP-13 inhibitors using a combined approach of machine learning, docking, and molecular dynamics

Abdul Manan¹ · Sidra Ilyas² · Eunha Kim¹ · Sangdun Choi^{1,3} · Donghun Lee²

Received: 2 May 2025 / Accepted: 6 September 2025

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2025

Abstract

Matrix metalloproteinase-13 (MMP-13) is a zinc-dependent endopeptidase involved in extracellular matrix degradation and inflammation, contributing to the progression of various diseases. This study applied an integrated computational approach encompassing QSAR modeling, machine learning (ML), scaffold analysis, docking, and molecular dynamics (MD) simulations to investigate the structure–activity relationships and binding mechanisms of MMP-13 inhibitors. A curated dataset of 1,741 unique compounds from ChEMBL was used to develop predictive QSAR models based on PubChem fingerprints. Among eight regression models, LGBM, SVR, and RF exhibited superior predictive performance, with LGBM achieving the best generalization (test RMSE = 0.825, $R^2 = 0.646$, $Q^2 = 0.628$). Similarly, LGBM and SVM classifiers demonstrated high accuracy (0.802) and MCC (0.589) with test data. Docking analysis identified three top candidates (ChEMBL1770157, ChEMBL425020 and ChEMBL5182668) with strong binding affinities of -10.98 , -10.93 and -10.80 kcal/mol, respectively. The identified interaction hotspots, particularly Thr245, Ala186, Leu185, Val219, and the highly versatile His222, represent key residues to target for enhancing binding affinity. Subsequent 200 ns MD simulations confirmed their structural stability and favorable binding dynamics within the MMP-13 active site. Scaffold analysis revealed the predominance of sulfonamide and carboxyl-containing polar functional groups, known to be important for solubility and target binding. The findings underscore the importance of physicochemical and structural attributes in MMP-13 inhibitor design and support the therapeutic potential of targeting MMP-13 in diverse pathological contexts.

Abdul Manan and Sidra Ilyas have contributed equally to this work.

✉ Sidra Ilyas
sidrailys6@gachon.ac.kr

✉ Donghun Lee
dlee@gachon.ac.kr

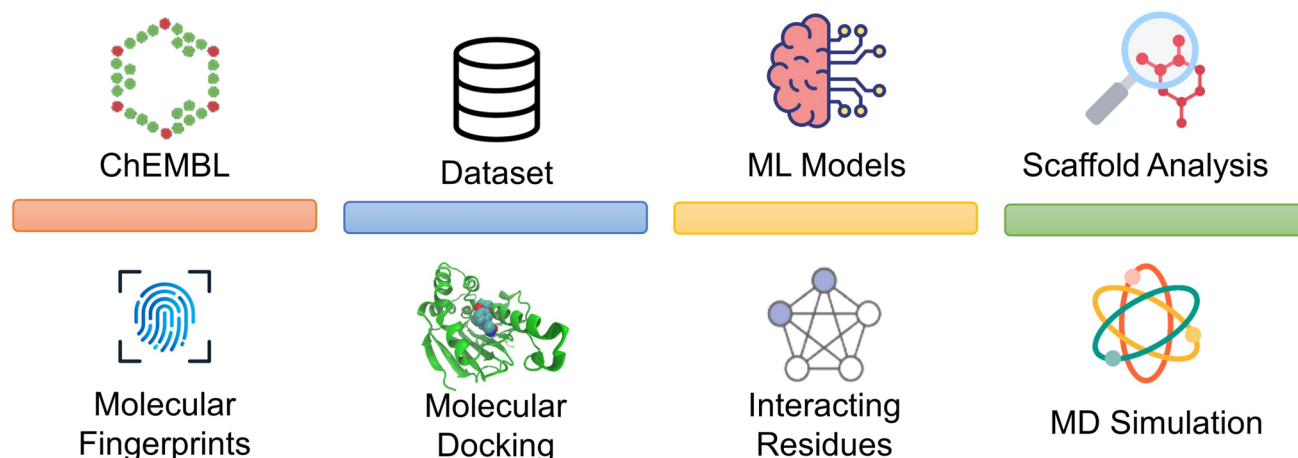
Abdul Manan
mananriaz012@gmail.com

¹ Department of Molecular Science and Technology, Ajou University, Suwon 16499, Korea

² Department of Herbal Pharmacology, College of Korean Medicine, Gachon University, 1342 Seongnamdaero, Sujeong-Gu, Seongnam-Si 13120, Korea

³ S&K Therapeutics, Suwon 16502, Korea

Graphical abstract



MMP-13 inhibitors targeting multiple diseases

Keywords MMP-13 · Machine learning · Docking · MD simulation · Bone diseases · Cancer · QSAR

Introduction

Matrix metalloproteinases (MMPs) are integral to extracellular matrix (ECM) homeostasis and involved in a wide range of physiological processes, including tissue morphogenesis, angiogenesis, immune regulation, and repair mechanisms [1]. ECM is a multifunctional, non-cellular 3D network composed of collagens, proteoglycans, elastin, fibronectin, laminins, and several glycoproteins. Within the MMPs family, MMP-13 (collagenase-3) is distinguished by its high catalytic efficiency and its pivotal role in cartilage and bone matrix turnover. MMP-13 is a zinc-dependent endopeptidase that plays a critical role in the degradation and remodeling of ECM. It exhibits a broad substrate specificity, cleaving various ECM components, including fibrillar collagens (types I, II and III), with the highest activity observed against soluble type II collagen, the primary structural protein in articular cartilage and intervertebral discs [2]. MMP-13 also degrades collagen (types IV, X and XIV), fibronectin, tenascin-C (TNC), and aggrecan (ACAN) by activating/degrading the key regulatory proteins, including transforming growth factor-beta 1 (TGFB1) and connective tissue growth factor (CCN2), contributing to its multifaceted role in tissue dynamics [3, 4].

Physiologically, MMP-13 is essential for normal embryonic bone development, endochondral ossification, bone mineralization, and bone healing [5]. It also facilitates wound healing and keratinocyte migration through the activation of TGFB1 and degradation of CCN2 [6]. MMP-13

expression is tightly regulated by factors such as parathyroid hormone-related protein (PTHrP) and the transcription factor Runx2 in chondrocytes, emphasizing its importance during skeletal development and endochondral ossification [7, 8]. However, dysregulation of MMP-13 activity is implicated in various pathological conditions such as osteoarthritis (OA), rheumatoid arthritis (RA), osteoporosis (OP) and cancer, where excessive ECM degradation leads to tissue destruction and disease progression [5, 9].

In OA, a degenerative joint disorder, MMP-13 drives cartilage degradation by breaking down type II collagen [10]. This enzymatic activity is exacerbated by aging, obesity, mechanical stress, and inflammation, which stimulate IL-1 β , TNF- α , NF- κ B, and MAPK signaling pathways [11]. These cascades not only promote catabolic activity but also generate ECM fragments that perpetuate synovial inflammation, creating a vicious cycle of tissue degradation. Elevated MMP-13 levels in cartilage and serum of OA patients correlate with disease severity and have been shown to induce OA-like joint pathology in animal models [8, 12]. RA, a chronic autoimmune disorder, is marked by persistent synovial inflammation and progressive joint destruction. MMP-13 is overexpressed in synovial fibroblasts and chondrocytes, facilitating cartilage degradation and bone erosion in RA [13]. This process is driven by inflammatory mediators such as IL-6, IL-17, and TNF- α through JAK/STAT and related pathways [14]. MMP-13, upregulated by cytokines such as IL-1 β , further contributes to ECM breakdown and cartilage damage, with elevated serum levels

observed in RA patients. In OP, MMP-13 promotes bone matrix degradation, facilitating osteoclastic bone resorption through pathways involving RANKL and TGF- β [15].

MMP-13 also plays a crucial role in tumor progression by degrading ECM barriers, promoting epithelial-mesenchymal transition (EMT), and interacting with the tumor microenvironment through factors such as VEGF, TAMs, thus activating Akt, STAT3, and EGFR signaling pathways [16]. Interestingly, its role can be context-dependent, influencing ECM stiffness and immune interactions, particularly in breast cancer [17]. In IVDD, a major contributor to chronic back pain, MMP-13 is significantly upregulated in degenerated discs, especially in the nucleus pulposus, where it compromises ECM integrity. Pro-inflammatory cytokines (TNF- α , IL-1 β , IL-6, IL-17) further elevate MMP-13 expression, linking inflammation to structural degradation [18]. Promising therapeutic interventions, such as BRD4 inhibition and anti-inflammatory agents (dimethyl fumarate, gastrodin), have shown potential in reducing MMP-13 levels and preserving disc function [19]. Additionally, increased MMP-13 levels in infectious disc diseases such as pyogenic spondylitis highlight its broader pathological relevance.

Given its widespread involvement in multiple diseases and correlation with disease severity, MMP-13 is a promising biomarker and therapeutic target [20, 21]. Although candidates such as licofelone, CDDO (triterpenoid), cipemastat, and ilomastat have shown potential, concerns about toxicity and suboptimal efficacy persist [22–26]. However, current MMP inhibitors often lack selectivity and produce off-target effects, limiting their clinical utility. Thus, there is an urgent need to develop highly selective and low-toxicity MMP-13 inhibitors that can effectively modulate its activity while preserving normal ECM turnover. To this end, computational approaches such as Quantitative Structure–Activity Relationship (QSAR) modeling offer valuable tools for drug discovery. QSAR models link the structural and physicochemical properties of compounds with their biological activities, enabling the rational design of effective inhibitors. Modern QSAR approaches practice advanced machine learning (ML) algorithms and robust validation techniques to explore chemical space with high accuracy. These models follow guidelines established by the Organization for Economic Cooperation and Development (OECD), which include defined endpoints, clear methodologies, applicability domain (AD), validation, and mechanistic interpretation [27].

Over the past decades, QSAR studies have played a significant role in the rational design of MMP-13 inhibitors [28]. Early QSAR models aimed to elucidate the physicochemical properties and structural features crucial for inhibitory activity against MMP-13 by correlating molecular descriptors with biological activity. These

studies often explored various classes of compounds, such as anthranilic acid and N-hydroxy-phenylsulfonylecetamide derivatives, identifying key descriptors related to atomic properties, electronic effects, and structural fragments [29]. Recent developments have seen the application of advanced chemometric methods, including combinatorial protocol in multiple linear regression (CP-MLR) and partial least-squares (PLS) analysis, to build more robust and predictive QSAR models [30]. These advanced approaches have enabled a deeper understanding of the structure–activity relationships, particularly for novel scaffolds like quinazolinone and fused pyrimidine derivatives, and have facilitated the design of potent and selective MMP-13 inhibitors with improved pharmacokinetic properties, thereby guiding targeted drug discovery efforts [31]. Building upon these advancements, the present study contributes to the field by uniquely combining a comprehensive QSAR modeling and docking analysis of a large dataset of known zinc-binding MMP-13 inhibitors, allowing for a refined understanding of their specific interactions and contributions to binding mechanism within the MMP-13 active site.

This study presents an integrated computational strategy to predict the inhibitory bioactivity of MMP-13 inhibitors using QSAR modeling and machine learning (ML) techniques, based on bioactivity data retrieved from the ChEMBL database. Compounds were categorized as active or inactive according to their pIC₅₀ values, and their chemical space was explored through exploratory data analysis (EDA). QSAR models were developed following OECD guidelines [32]. A diverse range of molecular descriptors and ML algorithms was employed to ensure the robustness and interpretability of the models. Scaffold analysis was conducted to identify key structural motifs contributing to MMP-13 bioactivity. By incorporating QSAR modeling, scaffold analysis, ML, docking, and molecular dynamics (MD) simulations, this research aims to optimize lead compounds targeting MMP-13. This integrated pipeline facilitates the rational design of novel MMP-13 inhibitors for the potential treatment of several MMP-13 associated diseases, including RA, OA, OP, cancer, and IVDD.

Experimental procedures

The workflow of the research is presented in Fig. 1. The methodology consisted of data collection, data pre-processing and cleansing, exploratory data analysis (EDA), quantitative structure–activity relationship (QSAR) modelling, scaffold analysis, docking, and MD simulation.

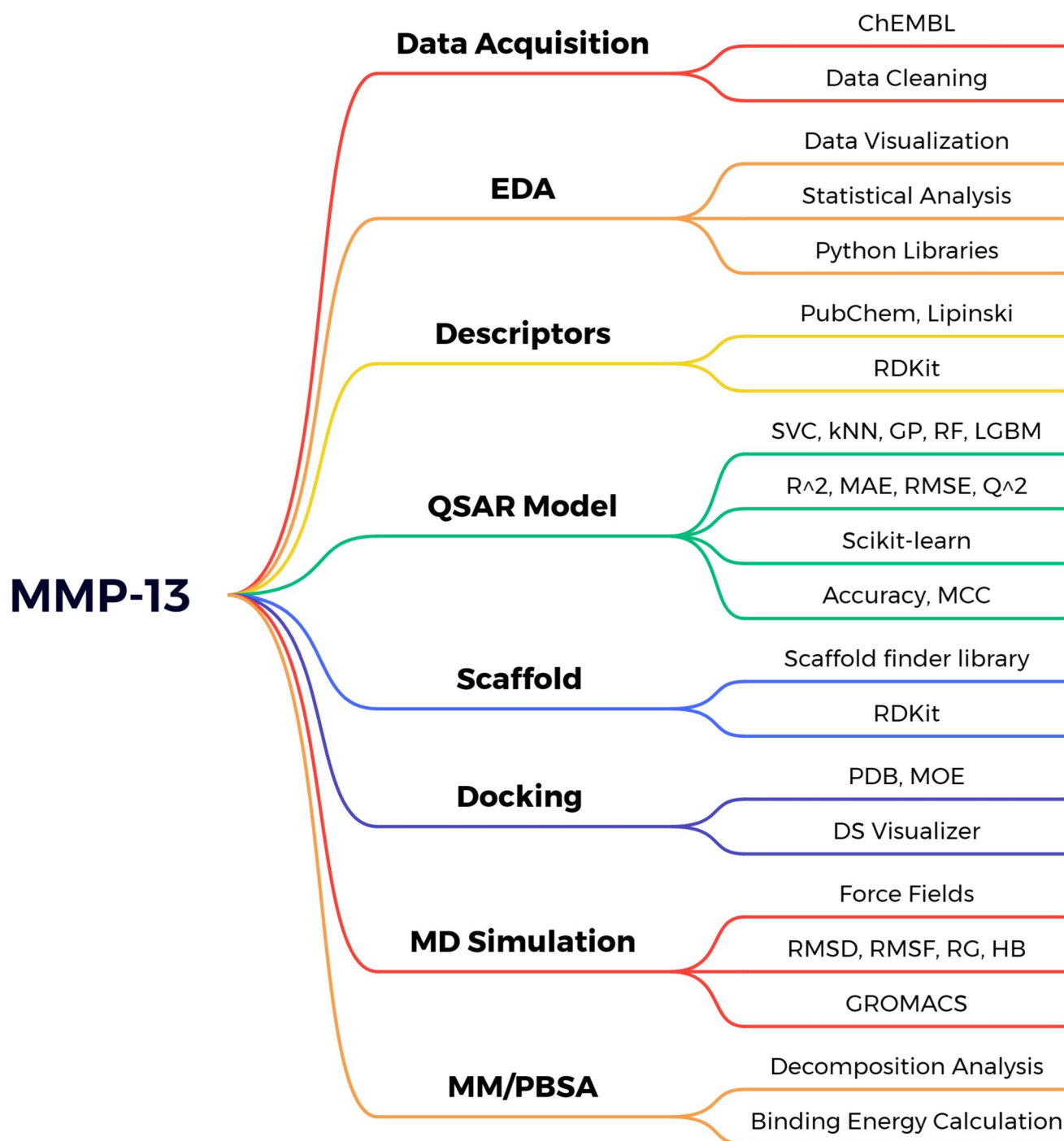


Fig. 1 The workflow used in the study

Data compilation

The bioactivity data of MMP-13 was retrieved from the ChEMBL database (v.31; target ID: 280) with a total of 3,191 points with an IC₅₀ value. Prior to working, the dataset was pre-processed and refined, which involved handling missing values, such as IC₅₀ (bioactivity measurement) and

SMILES (chemical structure representation) and removal by using the pandas library in Python. Duplicate entries in the ChEMBL dataset (same SMILES) were filtered using RDKit canonicalization. To further characterize the inhibitors and distinguish between zinc-binding (ZBI) and non-zinc-binding inhibitors (NZBI), compounds were analyzed based on their zinc-binding groups (ZBGs).

Common ZBGs were identified and categorized using specific SMARTS patterns as follows: hydroxamate ($C(=O)N(OH)$), carboxylate ($C(=O)(O;OH)$), thiol (SH), phosphonate ($P(=O)(O)O$), and sulfonamide ($S(=O)(=O)(NH)$). This categorization allowed for a more refined analysis of binding mechanisms and specific interactions with the catalytic zinc ion. The resulting dataset consists of 1,741 compounds. Next, the bioactivity activities (IC_{50}) were standardized by taking logarithmic pIC_{50} values using the formula $pIC_{50} = -\log_{10}(IC_{50} \text{ in Molar})$, as biological responses to drug concentrations typically follow a logarithmic pattern. The transformation was performed by applying the negative logarithm to the base of 10 to ensure consistency in data representation. Subsequently, inhibitors were classified into two categories: active ($pIC_{50} \geq 7.5$) and inactive ($pIC_{50} \leq 7.5$).

Exploratory data analysis (EDA)

Lipinski's rule of five is a widely used guideline for assessing the drug-like properties of molecules, aiding in the understanding of the structure–activity relationships [33]. To evaluate these properties, six physicochemical descriptors such as molecular weight (MW), octanol–water partition coefficient (LogP), number of hydrogen donors (nHD), number of hydrogen acceptors (nHA), number of rotatable bonds (nRot) and topological polar surface area (TPSA) were computed. Additionally, chemical space analysis, which represents the theoretical space where a molecule exists, is conducted to characterize the general properties of both active and inactive compounds.

Statistical analyses

To analyse the diverse trends and patterns of molecular descriptors in active and inactive compounds, a combination of statistical and exploratory data analysis techniques was employed. The nonparametric test Mann–Whitney U test was used to assess differences in the distribution of molecular descriptors between the two groups. Additionally, univariate statistical analysis of six descriptive key parameters such as mean, median, first quartile (Q1), third quartile (Q3), minimum (Min) and maximum (Max) was done. Molecular overlapping and distribution pattern of six parameters were explored using principal component analysis (PCA), which reduced the dimensionality of the dataset while retaining most of its information. Furthermore, PCA was also utilized for applicability domain (AD) visualization of the model; Local Outlier Factor (LOF) method was applied to identify potential outliers within the model [34].

Molecular descriptor generation and calculation

Molecular fingerprints serve as advanced representations of small molecules, capturing their structural features, interconnectivity, similarities, and physicochemical properties through bit strings comparison. PubChem fingerprint, consisting of 881 binary (1/0) bits, encodes the presence or absence of specific molecular fragments within a molecule. This fingerprinting method is implemented using the free and open-source PaDEL software package, which efficiently generates molecular descriptors for computational analysis [35]. Feature selection is performed to enhance the accuracy of the QSAR model, prevent overfitting, and eliminate bias, as all these factors negatively affect the results. A correlation-based filter method is applied to reduce feature complexity by removing highly correlated features (correlation > 0.90) and those with low variance (variance < 0.1) [36].

QSAR model building and validation

Independent QSAR classification and regression models were constructed to predict compound bioactivity. The algorithm yielding the best performance was selected following a thorough evaluation. Both internal and external validation procedures were performed to ensure model reliability. The dataset was partitioned into training and test sets, with a 4:1 ratio. To improve model generalization within the training set, a tenfold cross-validation strategy was implemented. This involved randomly dividing the training data into ten equal folds, using nine for model training and one for testing, repeated ten times with each fold serving as the test set once. The average accuracy across these iterations provided a measure of predictive performance. External validation, a critical step to assess the model's ability to accurately predict bioactivity for unseen compounds, was conducted using accuracy (ac), recall (re), and Matthew's correlation coefficient (MCC) as evaluation metrics [37].

The accuracy, precision, and MCC metrics were calculated using the following equations:

$$\text{Accuracy} = \frac{TP + TN}{(TP + FP + FN + TN)}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{MCC} = \frac{(TP \times TN - FP \times FN)}{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}$$

where TP, TN, FP and FN stand for true positive, true negative, false positive and false negative, respectively. The

accuracy and recall values range from 0 to 1 (worst to best) while the MCC value ranges from -1 to 1 (misclassification to perfect classification).

Standard regression metrics, including root mean square error (RMSE), coefficient of determination R-squared (R^2), mean absolute percentage error (MAPE), and cross-validated Q^2 were used. Root mean square error (RMSE) is a measure of standard deviation. R-squared (R^2) measures the proportion of the variance explained by the model. Average percentage of absolute errors (MAPE) measures the average percentage of the absolute errors. Q^2 assesses predictive performance on a test set, using unseen data.

The RMSE, R^2 , Q^2 , and MAPE were calculated using the following equations:

$$RMSE = \sqrt{\frac{1}{n} \sum (\hat{y}_i - y_i)^2}$$

$$R^2 = 1 - \frac{\sum (\hat{y}_i - y_i)^2}{\sum (y_i - \bar{y})^2}$$

$$Q^2 = 1 - \frac{\sum (\hat{y}_i^v - y_i)^2}{\sum (y_i - \bar{y})^2}$$

$$MAPE = \frac{100\%}{n} \sum \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

where y_i and $\hat{y}_i$ are the observed and predicted values, respectively.

Molecular docking

Docking was conducted to evaluate the binding interactions between ligands and MMP-13 protein (PDB ID: 3KRY) using the Molecular Operating Environment (MOE) software, version 2020.09 [38]. The protein was optimized within MOE, which included protonation, energy minimization, and correction of any structural errors. The crystal structure of MMP-13 was prepared and optimized within MOE, where the active site was defined based on known ligand-binding regions. The active site was centered at the coordinates $X = -10.715086$, $Y = -0.709400$, and $Z = 1.258114$. The interaction with the catalytic zinc ion (Zn1999) was enforced as a crucial constraint during the docking process. A comprehensive dataset of 1,741 known active MMP-13 inhibitors was selected for this study. Each ligand was prepared within MOE 2020.09. This preparation involved generating 3D conformations, protonation state adjustments at physiological pH, and energy minimization to obtain their most stable geometries for accurate docking. Standard docking protocols

were applied, and the resulting protein–ligand complexes were analyzed to identify key molecular interactions such as hydrogen bonding, π – π stacking, and hydrophobic contacts that contribute to binding affinity and specificity. For detailed visualization of the docking poses and interaction profiles, Discovery Studio Visualizer (v21.1.0) was employed to generate 2D interaction diagrams. A dataset comprising 1,373 known active MMP-13 inhibitors was selected for docking.

MD simulation, analysis and visualization

MD simulations were performed to investigate the dynamic behavior of protein–ligand complexes. Initial docked complexes were parameterized using CHARMM36 all-atom force field. Each complex was solvated in a cubic box of TIP3P water molecules, maintaining a minimum solute-box edge distance of $12 \text{ \AA}$ [39]. Ligand topologies were generated via the CGenFF server (<https://cgenff.com/>) and adapted for use with GROMACS 2024.1. Sodium and chloride ions were added to neutralize the system. Energy minimization was conducted using the steepest descent algorithm, converging to a force tolerance of 0.001 kJ/mol . Following minimization, systems were equilibrated under isothermal-isobaric (NPT) conditions using the Nose–Hoover thermostat and Parrinello–Rahman barostat [40]. Short-range electrostatic and van der Waals interactions were calculated with a 1.2 nm cutoff, while long-range electrostatic interactions were treated using the Particle mesh Ewald (PME) method [41]. Production MD simulations were conducted for 200 ns per complex using GROMACS 2024.1 (<https://www.gromacs.org/>). Trajectory analysis was performed using GROMACS utilities, and structural visualizations were generated to interpret protein–ligand complex dynamics [42].

MM/PBSA binding free energy of the ligand–protein complexes

To further assess the energetic contributions to ligand binding, Molecular Mechanics/Poisson–Boltzmann Surface Area (MM/PBSA) calculations were performed using representative snapshots extracted from molecular dynamics (MD) trajectories. The binding free energy (ΔG_{bind}) is calculated and decomposed into distinct energetic components: van der Waals interactions (ΔE_{vdW}), electrostatic interactions (ΔE_{elec}), polar solvation energy (ΔG_{pol}), and nonpolar solvation energy (ΔG_{npolar}).

The total binding free energy is computed using the following equation:

$$\Delta G_{\text{bind}} = \Delta E_{\text{vdW}} + \Delta E_{\text{elec}} + \Delta G_{\text{solv}} - T\Delta S$$

$$\text{where } \Delta G_{\text{solv}} = \Delta G_{\text{pol}} + \Delta G_{\text{npolar}}$$

Additionally, per-residue free energy decomposition analysis was conducted to pinpoint amino acid residues that contribute significantly to ligand recognition and stabilization. This was achieved by calculating the energy contributions of individual residues in the bound state and comparing them to the unbound conformations. Residues exhibiting substantial differences in energy contributions were identified as critical for ligand binding. This integrated approach provided valuable insights into the molecular determinants of ligand–protein interactions and highlighted potential hotspots for future drug design efforts.

Scaffold identification

The methodology by Naveja et al. [43] was employed for the identification of common scaffolds within this dataset [43]. This approach involves fragmenting each molecule, filtering fragments based on size, and grouping identical fragments to identify prevalent scaffolds. The dataset of 2,593 molecules, represented as SMILES strings, was processed to identify frequent molecular scaffolds. Scaffold identification was performed using the *find_scaffolds* function from the *scaffold_finder* library, which decomposed molecules into fragments, filtered out smaller fragments, and grouped structurally similar fragments into scaffolds. The frequency of each scaffold's occurrence was recorded. The most frequently occurring scaffolds were then selected for further analysis using the *get_molecules_with_scaffold* function, which also identified R-group substitution points. Molecules sharing a common scaffold were aligned using RDKit's *GenerateDepictionMatching2DStructure* function to facilitate structural comparison. Finally, the aligned molecules were visualized with corresponding pIC_{50} values using the *mols2grid* library, enabling the identification of SAR trends within the dataset.

Results and discussion

Dataset preparation and classification

In this study, a dataset of MMP-13 inhibitors was curated from the ChEMBL database (v.31; Target ID: 280), yielding 3,191 compounds based on IC_{50} values. After data cleaning and pre-processing to remove duplicates and non-standardized entries, a final dataset consisting of 1,741 unique molecules (SMILES) was retained for analysis. IC_{50} values were converted into their negative logarithmic form (pIC_{50}) to standardize the measurement scale and enable comparison across compounds. Molecules with $\text{pIC}_{50} \geq 7.5$ were categorized as “active” while those below this threshold were considered “inactive” (Fig. 2A). The majority of compounds exhibited molecular weights within

the acceptable drug-likeness range (typically < 500–700 Da), indicating their potential pharmacokinetic favourability.

Exploratory data analysis (EDA)

To assess physicochemical determinants of bioactivity, six key molecular descriptors were analysed: molecular weight (MW), lipophilicity (LogP), number of hydrogen bond acceptors (nHA), hydrogen bond donors (nHD), number of rotatable bonds (nRot), and topological polar surface area (TPSA) (Table 1, Fig. 2B). The Mann–Whitney U test revealed statistically significant differences ($p < 0.05$) across all six descriptors between active and inactive groups, indicating that these features influence MMP-13 inhibitory activity.

The average MW of active compounds (467.11 Da) was slightly greater than that of inactive (447.70 Da), with a p-value of 1.28×10^{-7} . The higher molecular complexity may contribute to enhanced interactions with the enzyme's active site. While increased MW may limit permeability, it can also support more extensive binding interactions, which is particularly advantageous for large protein interfaces. Active compounds exhibited slightly lower mean LogP values (2.59) compared to inactive (2.62), with a statistically significant p-value of 6.77×10^{-1} . Excessive LogP values (up to 7.52) in some inactive compounds may reduce aqueous solubility and promote nonspecific interactions, potentially compromising target specificity and pharmacokinetics (Fig. 2B).

A significant enrichment of hydrogen bond acceptors was observed in active compounds (mean = 6.42) relative to inactive compounds (mean = 5.94), with a p-value of 4.64×10^{-12} . A higher number of acceptors may enhance interaction strength and selectivity through specific polar contacts within the binding site, while still maintaining a favourable permeability profile. Active molecules also contained slightly more hydrogen bond donors (mean = 2.36) than inactive compounds (mean = 2.29), supported by a p-value of 3.15×10^{-2} . Hydrogen bonding donors enhance the formation of stable, directional interactions with target residues, contributing to increased affinity. Flexibility, as indicated by the number of rotatable bonds (nRot), was higher in active compounds than inactive compounds. While flexible molecules can better conform to binding site geometries, excessive flexibility may incur entropic penalties. The active compounds seem to maintain a balance between adaptability and rigidity, supporting efficient binding. TPSA emerged as the most discriminating descriptor, with a mean value of $115.87 \text{ \AA}^2$ for active compounds versus $110.96 \text{ \AA}^2$ for inactive (p-value 3.17×10^{-6}). TPSA correlates with hydrogen bonding potential and compound polarity, both of which are critical for bioavailability and molecular recognition (Fig. 3).

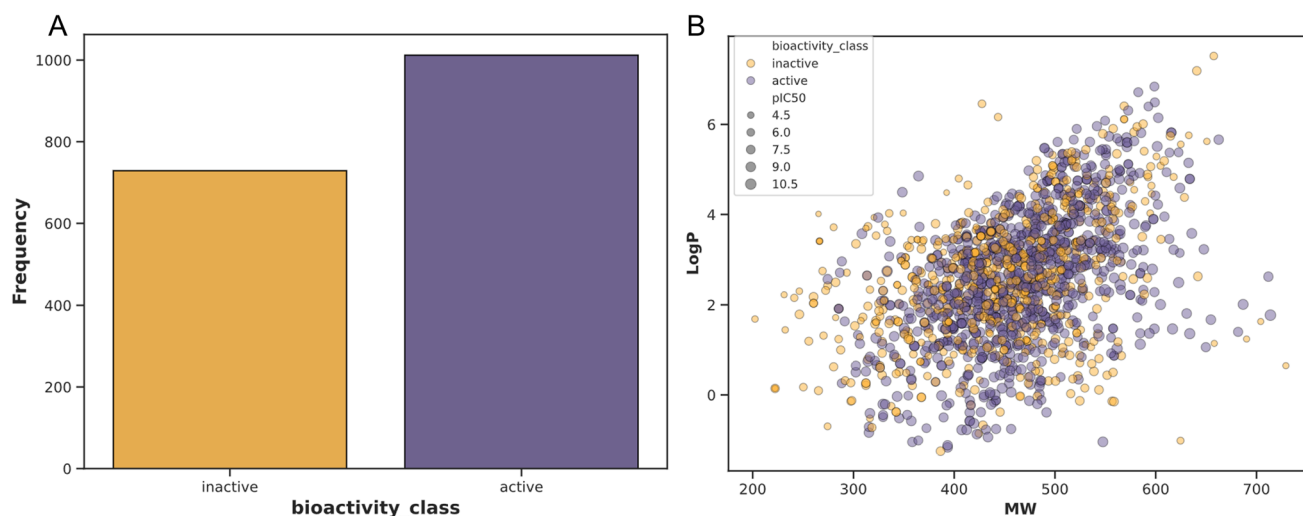


Fig. 2 The frequency and chemical space visualizations. **A** Frequency distribution of MMP-13 inhibitors. **B** Chemical space visualization based on a scatter plot of molecular weight (MW) versus octanol–

water partition coefficient (LogP), illustrating the distribution of active (purple), inactive (orange) compounds

Table 1 Summary of physicochemical descriptors of active (A1) and inactive (A2) compounds

Descriptor	A1 MW	A2	A1 LogP	A2	A1 nHA	A2	A1 nHD	A2	A1 nRot	A2	A1 TPSA	A2
p-value	1.28×10^{-7}		6.77×10^{-1}		4.64×10^{-12}		3.15×10^{-2}		4.56×10^{-4}		3.17×10^{-6}	
Min	274.41	202.25	−1.16	−1.25	3	2	0	0	3	2	34.14	42.09
Max	713.83	729.18	6.84	7.52	14	15	7	9	27	18	203.1	245.29
Median	465.50	452.55	2.60	2.60	6	6	2	2	7	7	116.17	111.16
Mean	467.11	447.70	2.59	2.62	6.42	5.94	2.36	2.29	7.77	7.34	115.87	110.96
Skew	0.11	−0.20	0.00	0.12	0.87	0.67	1.29	1.48	2.28	0.78	0.32	0.70
Kurtosis	0.15	0.26	−0.15	0.06	3.75	2.71	4.25	7.56	9.99	2.34	2.21	4.41

Where MW: molecular weight, LogP: octanol–water partition coefficient, nHA: No. of hydrogen bond acceptors, nHD: No. of hydrogen bond donors, nRot: No. of rotatable bonds, TPSA: topological polar surface area

Collectively, these findings highlight that active MMP-13 inhibitors tend to exhibit complex polar interactions by exhibiting higher MW and TPSA, moderate lipophilicity, within the ideal range for drug-likeness. Greater hydrogen bonding potential enhances target affinity and slightly higher conformational flexibility enables better fit within dynamic binding pockets. Chemical space visualization and the MW vs. LogP scatter plot (Fig. 2B) further elucidate these trends. These patterns underscore how fine-tuning physicochemical properties can enhance potency.

Principal component analysis (PCA)

To gain a comprehensive understanding of the relationships between key physicochemical properties and the biological activity of MMP-13 inhibitors, Principal component analysis (PCA) is done. This dimensionality reduction technique facilitates the identification of dominant variance patterns

and allows for the visualization of compound clustering based on shared chemical features. PCA is performed on six physicochemical descriptors, and the loading matrix (Table 2) reveals the contributions of each descriptor to the first three principal components (PCs), which cumulatively account for 88% of the total variance in the dataset.

PC1 accounted for the highest proportion of variance and was primarily influenced by TPSA (0.52), MW (0.51), nHA (0.48), and nRot (0.43). These features are associated with molecular size, polarity, flexibility, and hydrogen bonding potential. Thus, PC1 represents a polarity-complexity, where compounds with higher PC1 scores are generally larger, more polar, and structurally more flexible. Notably, active compounds tended to cluster at higher PC1 values, consistent with earlier findings indicating elevated TPSA, nHA, and MW. PC2 was predominantly defined by LogP (0.71), MW (0.32), and nRot (0.30), reflecting lipophilicity and conformational flexibility. This component captures

variations relevant to membrane permeability and oral bioavailability. Interestingly, a strong negative contribution from nHD (-0.40) suggests that an increase in hydrogen bond donors may inversely correlate with lipophilicity. Compounds with high LogP but fewer H-bond donors and lower TPSA score higher on this axis, capturing the hydrophobicity-hydrophilicity balance, which is critical for bioavailability and membrane permeability. This component may help distinguish molecules optimized for passive permeability versus those that are too polar or too hydrophobic, affecting activity and drug-likeness. PC3 is dominated by nHD (0.79) and counterweighted by nHA (-0.47), indicating this component represents the relative contribution of hydrogen bond donors versus acceptors. This axis could reflect differences in interaction modes within the MMP-13 binding site, particularly since hydrogen bonding is crucial for anchoring inhibitors to key residues (e.g., His222, Thr245, and Phe252). Compounds scoring high on PC3 may rely more on donor interactions, possibly influencing binding orientation and affinity.

A biplot of PC1 and PC2 illustrates the distribution of active and inactive compounds in the PCA space (Fig. 4). While there is partial overlap between the two classes, a discernible trend emerges: active compounds are more densely clustered toward the right side of the PC1 axis and, to a lesser extent, the upper-right quadrant of the PCA plot. This aligns with the observed tendency for active

compounds to exhibit higher polarity, hydrogen bonding capacity, and molecular complexity. Inactive compounds, by contrast, are more broadly scattered along the PC2 axis, consistent with the wider variability in lipophilicity and flexibility parameters seen in boxplot analyses (Fig. 3).

QSAR modelling and performance evaluation of classification models

To predict the bioactivity of MMP-13 inhibitors, PubChem fingerprints were employed to represent molecular structures. A total of 64 informative features were retained after variance and correlation-based filtering (thresholds:

Table 2 PCA loading matrix of the six physicochemical properties across the first three principal components

Property	PC1	PC2	PC3
MW	0.51	0.32	0.00
LogP	0.07	0.71	0.37
nHD	0.22	-0.40	0.79
nHA	0.48	-0.18	-0.47
nRot	0.43	0.30	-0.03
TPSA	0.52	-0.32	0.09
Cumulative variance (%)	0.46	0.72	0.88

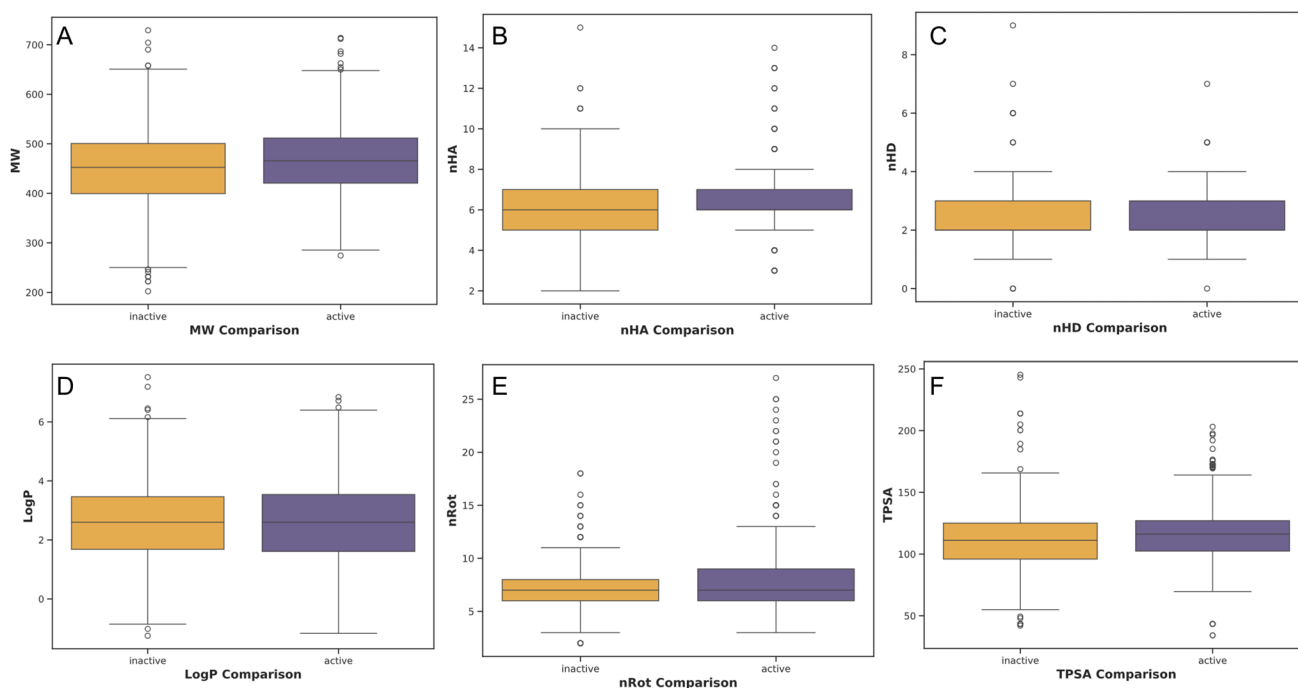


Fig. 3 Box plots comparing drug-likeness-related physicochemical properties between active and inactive compounds. Panels A–F show the distribution of molecular weight (MW), number of hydrogen bond

acceptors (nHA), number of hydrogen bond donors (nHD), LogP, number of rotatable bonds (nRot), and topological polar surface area (TPSA), respectively

0.10 and 0.95, respectively). Nine distinct classification and eight regression models were developed and optimized through hyperparameter tuning (supplementary table S1), with a fixed random state [42] to ensure reproducibility.

Among all tested models, Light Gradient Boosting Machine (LGBM) and Support Vector Classifier (SVC) exhibited the best generalization ability on the test set, both achieving an accuracy of 0.802 and MCC of 0.589. This highlights their robustness in correctly distinguishing between active and inactive compounds. LGBM also demonstrated excellent training performance (accuracy = 0.960, MCC = 0.917), suggesting that it effectively captures key patterns in the data without significant overfitting. Other ensemble methods such as Gradient Boosting (GB), Extra Trees (ET), and Random Forest (RF) also yielded strong results, with test accuracies ranging from 0.788 to 0.797 and MCC values between 0.559 and 0.578, showing reliable prediction and good consistency across splits. These tree-based methods are well-suited for handling high-dimensional molecular fingerprints and demonstrate stability in learning from complex structure–activity relationships.

The k-Nearest Neighbors (k-NN) and Gaussian Process (GP) classifiers achieved competitive training scores (accuracy = 0.960), but showed a slight drop in generalization performance on the test set (accuracy ~ 0.774–0.777; MCC ~ 0.532–0.536), suggesting a potential sensitivity to local noise and overfitting to training data. In contrast, Logistic Regression (LR) consistently underperformed,

with a test accuracy of 0.736 and the lowest MCC (0.448), indicating that its linear decision boundaries may be inadequate for capturing the non-linear relationships present in the dataset.

A comparative overview of all classifiers is summarized in Table 3. Performance is evaluated using tenfold cross-validation and independent test set validation. The results underscore the importance of using ensemble and kernel-based methods (e.g., LGBM, SVC, RF) for molecular classification tasks involving complex and non-linear chemical descriptors. LGBM, in particular, emerged as the most balanced model in terms of high accuracy and MCC across all data splits, suggesting its strong utility in virtual screening pipelines for MMP-13 inhibitor identification. The poor performance of LR supports prior evidence that linear models often fall short in bioactivity prediction due to the intricate nature of chemical space. Overall, these classification models offer effective tools for prioritizing compound libraries by identifying likely active candidates, thereby accelerating early stage drug discovery for MMP-13-related disorders.

Performance analysis of regression models

To evaluate the predictive performance of machine learning models for estimating the continuous inhibitory activity of MMP-13 compounds, eight regression algorithms were implemented and assessed using standard performance metrics: root mean square error (RMSE), coefficient

Fig. 4 Principal Component Analysis (PCA) of chemical space occupied by active (purple circles) and inactive (orange circles) compounds. The PCA is performed on all compounds regardless of their bioactivity to assess diversity in the dataset

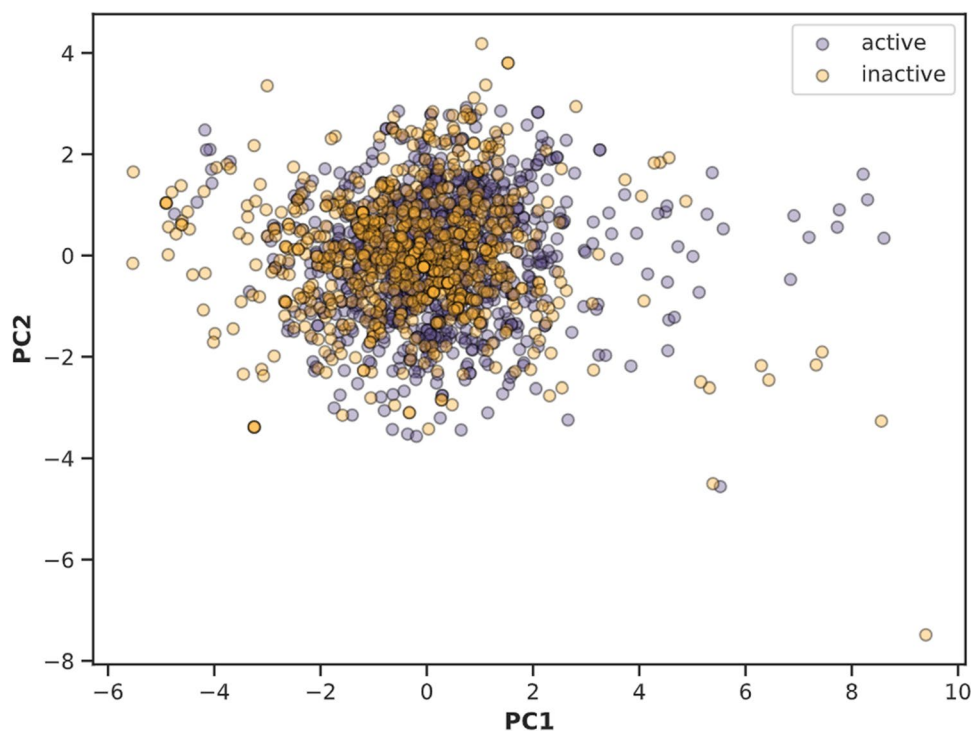


Table 3 Performance metrics for 9 classification algorithms. Structures of molecules represented by PubChem fingerprints with variance threshold (0.10), correlation threshold (0.95), random state (42)

Classification Models	Accuracy (acc)			MCC		
	Train	CV	Test	Train	CV	Test
DT	0.939	0.787	0.745	0.876	0.558	0.469
ET	0.960	0.812	0.797	0.917	0.611	0.578
RF	0.958	0.813	0.788	0.913	0.612	0.559
GB	0.960	0.819	0.788	0.917	0.627	0.559
LGBM	0.960	0.815	0.802	0.917	0.619	0.589
SVC	0.894	0.815	0.802	0.781	0.617	0.589
k-NN	0.960	0.810	0.777	0.919	0.607	0.536
LR	0.779	0.754	0.736	0.541	0.489	0.448
GP	0.960	0.811	0.774	0.918	0.609	0.532

Note: The model's performance is assessed by the accuracy (acc) and Matthew's correlation coefficient (MCC). Additionally, the performance is calculated separately for the training set (Train) with a tenfold cross-validation (CV) and the test set (Test)

of determination (R^2), mean absolute percentage error (MAPE), and cross-validated predictive squared correlation coefficient (Q^2 -CV). Among all models, LGBM and Support Vector Regression (SVR) showed the most favorable balance between training and test performance, making them the most robust and generalizable regressors in this study. LGBM achieved a train RMSE of 0.480, train R^2 of 0.879, and a low train MAPE of 4.66%. On the test set, it maintained strong performance with a test RMSE of 0.825, R^2 of 0.646, and MAPE of 8.51%, along with the highest Q^2 -CV (0.628) among all models.

SVR closely followed with the lowest train RMSE (0.46) and train MAPE (3.52%), indicating an excellent model fit. Its test metrics also remained competitive (RMSE=0.844, R^2 =0.630, MAPE=8.56%), confirming its stability and predictive accuracy. Random Forest (RF) also performed reliably, with a test RMSE of 0.824, R^2 =0.647, and MAPE=8.68%, and a Q^2 -CV of 0.627, reinforcing its effectiveness in handling non-linear relationships in molecular data. In contrast, models such as Decision Tree (DT) and Extra Trees (ET) showed signs of overfitting. Although they exhibited relatively low training errors (train RMSE: DT=0.561, ET=0.541), their test RMSEs increased significantly (DT=1.106, ET=1.068) and R^2 values dropped sharply (DT=0.364, ET=0.406), suggesting limited generalization capability. The higher test MAPE values (> 10%) further emphasize the inaccuracy of these models in external prediction. Gradient Boosting (GB) and k-Nearest Neighbors (k-NN) delivered moderate performance, with test RMSEs in the 0.84–0.85 range and Q^2 -CV values between 0.535–0.594, indicating that while they are capable of capturing trends, they are less reliable than SVR or LGBM for precise prediction. The Gaussian Process (GP) regressor achieved outstanding training performance (train RMSE=0.409, R^2 =0.912, MAPE=2.50%), but failed catastrophically in test prediction, with a test RMSE=5.714,

R^2 = −15.995, and MAPE=57.66%, indicating severe overfitting. Its negative Q^2 -CV (−16.509) confirms the model's inability to generalize and renders it unsuitable for regression in this context.

The comparative analysis reveals that LGBM, SVR, and RF are the most reliable regression models for predicting the bioactivity of MMP-13 inhibitors (Table 4). These models effectively balance learning capacity and generalization, making them ideal for quantitative structure–activity relationship (QSAR) modeling and virtual screening applications. Their ability to handle high-dimensional molecular descriptors while maintaining low prediction errors underscores their suitability for cheminformatics workflows. In contrast, models with high complexity but poor regularization (e.g., GP) or those prone to overfitting (e.g., DT, ET) show limited utility in predictive modeling despite good training performance. The findings align with current literature that supports the use of ensemble and kernel-based regressors for robust QSAR applications. These insights contribute to the rational selection of ML models in drug discovery pipelines and reinforce the importance of cross-validation and external test sets in evaluating model robustness.

Classification vs. regression models and applicability domain

Classification models such as LGBM and SVC achieved high test accuracies (0.80) and strong MCC values (0.58), making them effective for distinguishing active from inactive compounds. In regression tasks, LGBM, SVR, and RF again performed best, delivering low test RMSEs (~0.82–0.84), high R^2 (~0.63–0.65), and low MAPE values (< 9%), enabling reliable prediction of continuous bioactivity values. While GP regression showed excellent training performance,

Table 4 Performance metrics for 8 regression algorithms. Structures of molecules represented by PubChem fingerprints with variance threshold (0.10), correlation threshold (0.95), random state (42)

Regression Models	Train RMSE	Train R ²	Train MAPE	Test RMSE	Test R ²	Test MAPE	Q ² -CV
DT	0.561	0.834	5.310	1.106	0.364	10.524	0.440
ET	0.541	0.846	4.780	1.068	0.406	10.404	0.432
RF	0.518	0.859	5.302	0.824	0.647	8.683	0.627
GB	0.641	0.784	6.734	0.851	0.623	9.007	0.594
LGBM	0.480	0.879	4.663	0.825	0.646	8.509	0.628
SVR	0.460	0.889	3.518	0.844	0.630	8.556	0.618
k-NN	0.649	0.779	6.487	0.846	0.628	8.849	0.535
GP	0.409	0.912	2.503	5.714	−15.995	57.665	−16.509

Note: Root mean squared error (RMSE) is a measure of standard deviation. R-squared (R²) measures the proportion of the variance in the actual values that is explained by the variance in the predicted values. Average percentage of absolute errors (MAPE), measures the average percentage of the absolute errors, indicating the multiplicative effect between predicted and actual outcomes. Q-squared (Q²): Similar to R² but assesses predictive performance on a test set, using unseen data

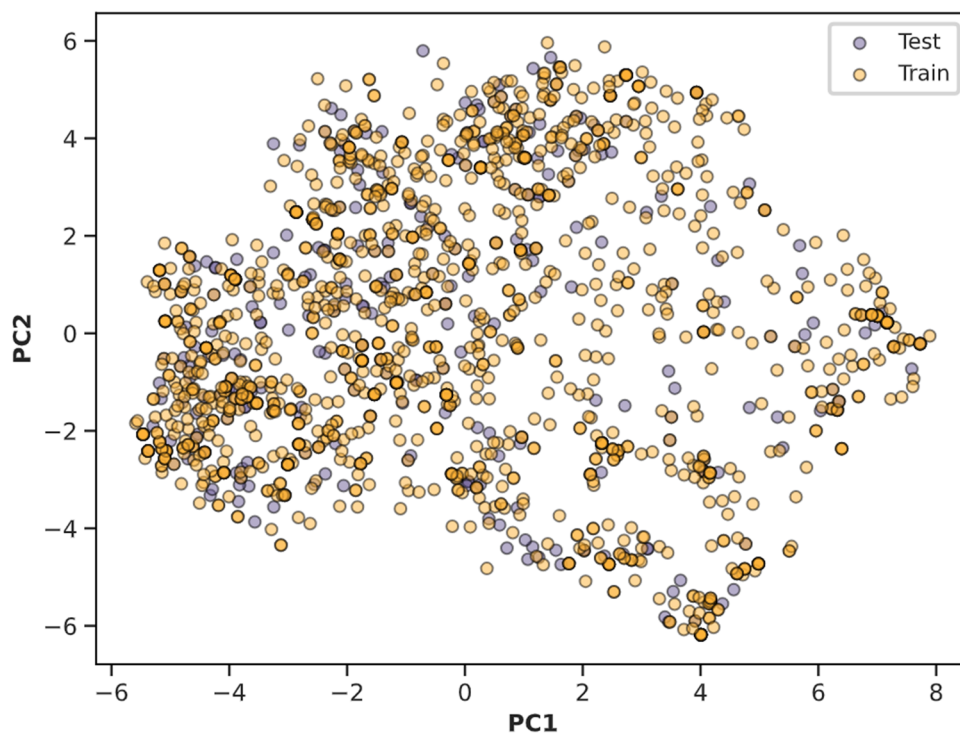
it failed to generalize, highlighting the risk of overfitting. Overall, LGBM and SVM emerged as top performers across both tasks, supporting their use in virtual screening pipelines. The findings emphasize that classification models are suitable for rapid activity filtering, whereas regression models are essential for potency estimation, and combining both provides a robust, complementary strategy in cheminformatics-driven drug discovery. Models like LGBM and SVM, which consistently perform well in both contexts, are especially suited for dual-role applications in drug discovery. Importantly, PCA-based applicability domain (AD) analysis confirmed that the test set lies within

the multidimensional chemical space of the training data (Fig. 5). This supports the reliability of predictions, aligning with OECD guidelines for QSAR modeling [27].

Scaffold identification and analysis

The analysis of 2,593 molecular entries identified 46,847 unique scaffolds. Notably, the most frequent scaffold 1, appeared 59 times, representing approximately 2.3% of the dataset. These frequent scaffolds suggest potential pharmacophoric features crucial for biological activity. Conversely, numerous scaffolds appeared only once,

Fig. 5 Applicability domain (AD) of the ML models visualized by using PCA. The training set (orange circles) and test set (purple circles) are projected in the chemical space, showing similar dispersion and confirming that the test compounds fall within the model's AD, thereby validating prediction reliability



highlighting the structural diversity of the dataset. Scaffold 1 comprises a piperidine ring, hydroxamic acid moiety, sulfonamide group, and a phenol ring, features commonly associated with bioactive compounds. Literature reports suggest that similar frameworks act as inhibitors of carbonic anhydrase (CA) [44]. The hydroxamic acid unit is a well-known zinc-chelating motif, relevant for MMP inhibition [45]. Additionally, the presence of piperidine and sulfonamide functionalities points to potential interaction with adrenergic receptors, notably β_3 agonists, which influence metabolic regulation [46]. Altogether, this scaffold presents multiple pharmacophores suited for enzyme inhibition and receptor modulation, relevant to oncology and inflammation.

Featuring a biphenyl sulfonamide linked to an acetic acid moiety, the scaffold 2 mirrors structures seen in known JAK inhibitors such as oclacitinib and carbonic anhydrase inhibitors [47]. Related derivatives have shown activity as CBKinasel inhibitors and angiotensin or endothelin receptor antagonists. The acidic and lipophilic balance provided by the sulfonamide and biphenyl units suggests its potential in targeting inflammation, cardiovascular, and neurodegenerative diseases. The scaffold 3 is structurally similar to N-(4-methoxyphenyl)glycine derivatives, and it shares functional elements with known cyclooxygenase-2

inhibitors [48]. The methoxy-substituted biphenyl moiety may modulate lipophilicity and electron distribution, possibly enhancing biological interactions. Its structural features support possible applications in enzyme inhibition and antiproliferative therapies.

Highly reminiscent of Matrix Metalloproteinase inhibitors, the scaffold 4 includes a methoxy phenyl sulfonamide and a hydroxamic acid group, both key to MMP binding [12]. Notably, the N-methyl sulfonamide enhances lipophilicity and may improve pharmacokinetics. Given its similarity to known MMP-9 inhibitors, this scaffold may hold value in cancer and inflammatory disease treatment strategies. A biphenyl sulfonamide bearing a cyano group, the scaffold 5 aligns with structures known for activity as receptor antagonists. The cyano substitution may act as a bioisostere, enhancing metabolic stability and modulating target affinity [49]. Literature supports potential applications in cardiovascular, anti-inflammatory, and anticancer domains. All five scaffolds share key pharmacophoric elements, particularly sulfonamides and aromatic systems that favor interactions with a variety of biological targets (Table 5). Biphenyl or biphenyl-like structures in Scaffolds 2, 3, and 5 are privileged motifs, commonly used across medicinal chemistry. Hydroxamic acids in Scaffolds 1 and 4 are indicative of MMP inhibition potential. Each scaffold

Table 5 Frequency distribution and number of atoms of the first five scaffold

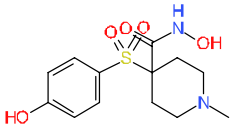
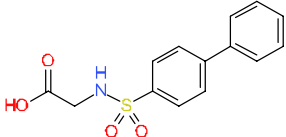
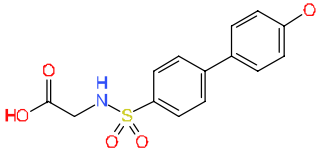
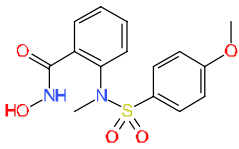
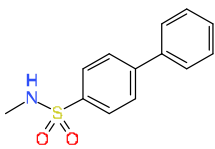
No	Frequency	Scaffold	Atom Number	Structures
1	59	<chem>CN1CCC(C(=O)NO)(S(=O)(=O)c2ccc(O)cc2)CC1</chem>	21	
2	59	<chem>O=C(O)CNS(=O)(=O)c1ccc(-c2ccccc2)cc1</chem>	20	
3	56	<chem>COc1ccc(-c2ccc(S(=O)(=O)NCC(=O)O)cc2)cc1</chem>	22	
4	51	<chem>COc1ccc(S(=O)(=O)N(C)c2ccccc2C(=O)NO)cc1</chem>	23	
5	51	<chem>CNS(=O)(=O)c1ccc(-c2ccccc2)cc1</chem>	17	

Table 6 The key functional groups, potential biological activities, and research areas for each scaffold

Scaffold No	Functional Groups	Potential Biological Activities	Research Areas	References
1	Piperidine, Hydroxamic acid, Sulfonamide, Phenol	CA inhibition, MMP inhibition, Beta-3 adrenergic receptor agonism, TNF- α inhibition	Medicinal chemistry, Drug discovery, Oncology, Inflammation, Metabolic disorders	[50]
2	Carboxylic acid, Secondary amine, Sulfonamide, Biphenyl	JAK inhibition, CA inhibition, CBKinase1 inhibition, Angiotensin/Endothelin receptor antagonism, PARP inhibition	Drug discovery, Inflammation, Allergy, Cardiovascular diseases, Glaucoma, Epilepsy, Neurodegenerative disorders	[51]
3	Methoxy, Biphenyl-like system, Sulfonamide, Secondary amine, Carboxylic acid	CA inhibition, Anticholinesterase inhibition, Urease inhibition, Antiproliferative activity	Drug discovery, Neurodegenerative disorders, Infectious diseases, Oncology	[52]
4	Methoxy, Sulfonamide, N-methyl, Phenyl, Hydroxamic acid	MMP inhibition	Drug discovery, Oncology, Inflammation	[53]
5	Cyano, Sulfonamide, Secondary amine, Biphenyl	Angiotensin/Endothelin receptor antagonism, enzyme inhibition	Drug discovery, Cardiovascular diseases, Inflammation, Oncology, Antimicrobial research	[49, 52]

also offers unique substituent patterns that may influence binding profiles and selectivity. The predominance of sulfonamide and carboxyl-containing scaffolds indicates a focus on polar functional groups, known to be important for solubility and target binding (Table 6).

Molecular docking studies

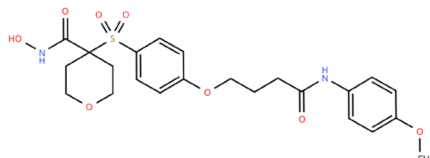
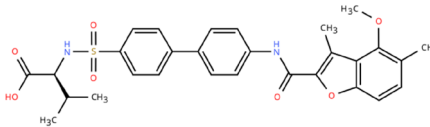
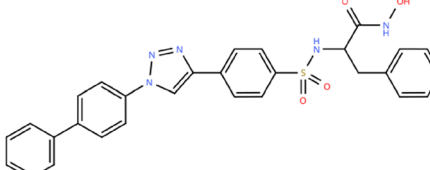
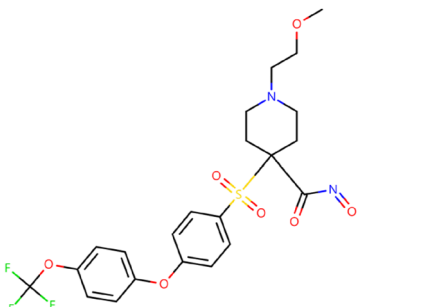
Docking analysis revealed three top-ranked candidate compounds, ChEMBL1770157, ChEMBL425020, and ChEMBL5182668, exhibiting favourable binding affinities of -10.98 , -10.93 , and -10.80 kcal/mol, respectively (Table 7). The reference molecule with binding score of -8.15 kcal/mol exhibits a multifaceted binding profile characterized by a strong anchoring through a π -anion interaction with Zn1999. It forms multiple conventional hydrogen bonds with Thr245, Ala186, and Leu185, indicating specific polar contacts. Additionally, carbon-hydrogen bonds are observed with Pro242, Val219, and His187, contributing to overall stability. A notable feature is the halogen bond with Thr245, suggesting that specific halogen atoms on the ligand can engage in precise interactions. Hydrophobic interactions include π -alkyl contacts with Leu239, Val219, and Leu185, and a π -sigma interaction with Tyr244 (Fig. 7A). Furthermore, a π - π T-shaped interaction with His222 highlights the engagement of aromatic moieties.

ChEMBL1770157 binding is also strongly characterized by a π -anion interaction with Zn1999. It forms hydrogen bonds with Thr245, Ala186, and Leu185, closely resembling the polar interactions of the reference. A unique carbon-hydrogen bond is observed with Gly237. Hydrophobic interactions are mediated through π -alkyl contacts with Pro255 and Val219, along with a π - π T-shaped interaction with His222 (Fig. 7B). ChEMBL425020 demonstrates a

complex interaction profile, featuring π -anion interactions with both Zn1999 and Glu223, suggesting potentially strong electrostatic attractions. It forms conventional hydrogen bonds with Thr245 and Ile243. Carbon-hydrogen bonds are noted with Tyr244 and Ala188. Hydrophobic interactions include π -alkyl contacts with His226, Val219, and Tyr176. Furthermore, ChEMBL425020 exhibits π - π T stacked interactions with His232 and His222, indicating significant aromatic stacking (Fig. 7C). Notably, an unfavorable donor-donor interaction is observed with His226, pointing to a potential energetic penalty or steric clash in this region. ChEMBL5182668 displays a robust set of interactions, also anchored by a π -anion interaction with Zn1999. Its polar contacts include conventional hydrogen bonds with Pro242 and Ala186 (Fig. 7D). A diverse array of carbon-hydrogen bonds is formed with Thr245, Phe241, His222, and Leu184, indicating extensive engagement within the binding pocket. The hydrophobic landscape is dominated by π -alkyl interactions with Pro255, Leu239, Val219, and Leu185. Significantly, ChEMBL5182668 engages in multiple π - π T-shaped interactions with Tyr244, His232, and His222, suggesting a strong reliance on aromatic stacking for binding stabilization (Fig. 6).

The comparative analysis reveals both conserved and unique binding strategies among the four molecules, providing a comprehensive understanding of the protein's ligand recognition capabilities (Fig. 7). While Zn1999 serves as a universal anchor, the specific hydrogen bonding and hydrophobic/aromatic interactions vary, displaying the binding pocket's adaptability. Thr245 consistently forms conventional hydrogen bonds across most molecules, and notably, a halogen bond with the reference. Ala186 and Leu185 also frequently participate in conventional hydrogen bonds. The hydrophobic/aromatic sub-pockets are highly utilized, with His222 emerging as a highly versatile

Table 7 The selected molecules from the potent category of MMP-13 inhibitors and their binding energies (kcal/mol) with IC₅₀ (nM) values are shown

ChEMBL ID	Binding Energy (kcal/mol)	IC ₅₀ (nM)	pIC ₅₀	Structures
1770157	− 10.98	0.3	9.52	
425020	− 10.93	4.9	8.30	
5182668	− 10.80	5	8.30	
3KRY Reference 49866490*	− 8.15	0.4 ^{\$}	9.4	

*PubChem ID; \$ = 10.1021/jm0500875 (DOI)

residue, engaging in various forms of π -interactions (π - π T-shaped, π -alkyl, π - π T Stacked) across all four molecules. Val219 and Leu239 are also critical for π -alkyl interactions. ChEMBL5182668 and ChEMBL425020 demonstrate a particular propensity for extensive aromatic stacking interactions involving multiple histidine and tyrosine residues. The presence of a halogen bond in the reference and the additional π -anion interaction with Glu223 in ChEMBL425020 highlights unique interaction strategies that contribute to their specific binding. The unfavorable interaction observed with His226 in ChEMBL425020, contrasting with the overall favorable binding, suggests a localized challenge that could be addressed through structural modification.

MD simulation studies

The RMSD plots for the protein–ligand complexes (ChEMBL425020, ChEMBL1770157, ChEMBL5182668 and 3KRY) over a 200 ns simulation provide insight into the structural stability of the systems. All ligand-bound complexes exhibit an initial rise in RMSD within the first 20–30 ns, indicative of structural relaxation and equilibration. Following this, the trajectories stabilize with RMSD values fluctuating between ~0.4–0.7 nm. ChEMBL1770157 (red) demonstrates the most consistent and stable RMSD profile after ~30 ns, maintaining values below 0.6 nm, suggesting a well-converged and stable binding (Fig. 8A). ChEMBL5182668 (green) and 3KRY (blue) also exhibit relatively stable behavior, though with moderate fluctuations. ChEMBL425020 (black) shows higher RMSD deviations throughout the simulation,

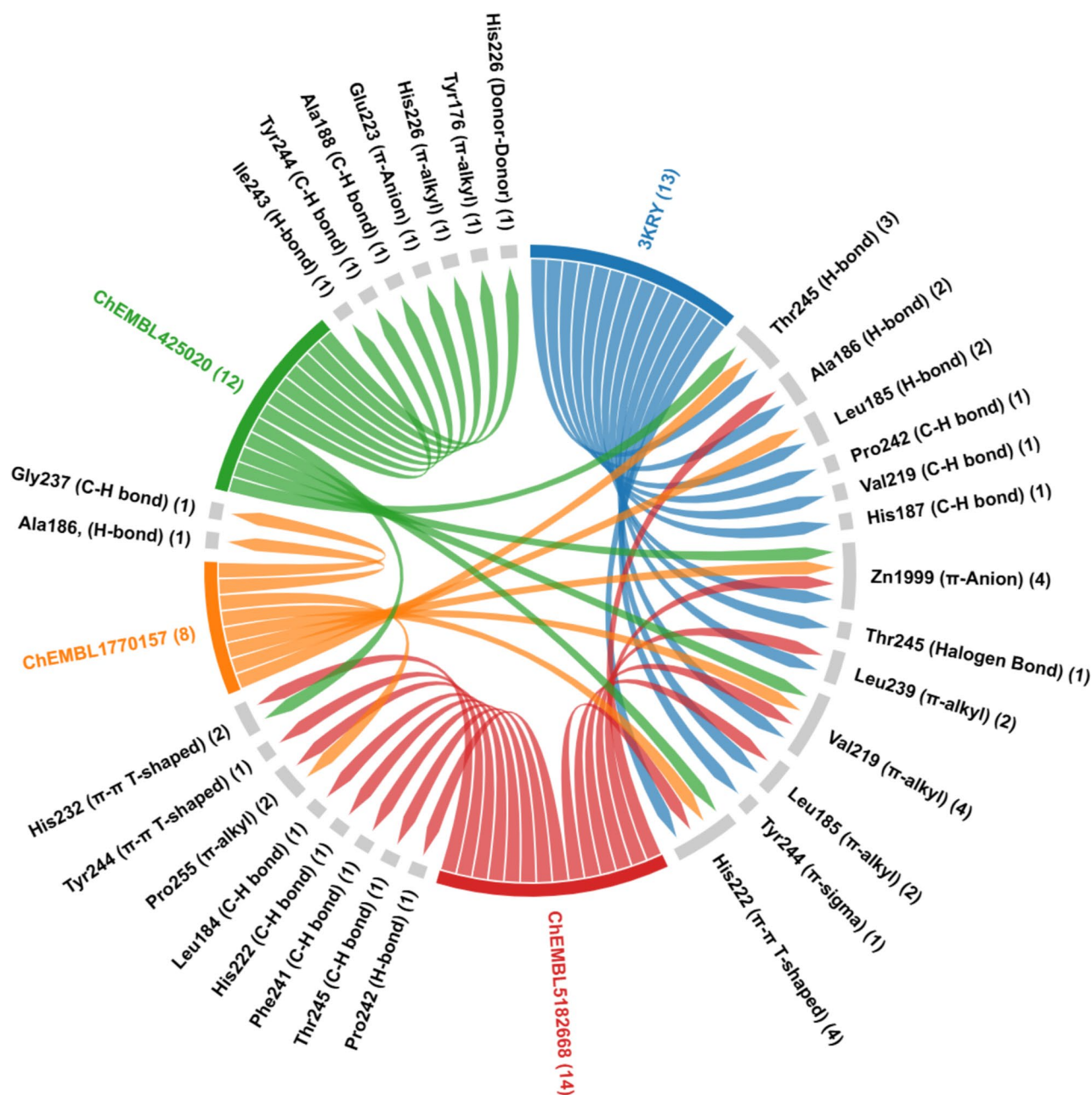


Fig. 6 Protein ligand interaction analysis of 3KRY, ChEMBL1770157, ChEMBL425020, and ChEMBL5182668. A chord diagram was used to illustrate the connectivity and distribution of interacting residues between the proteins and ligands. The diagram

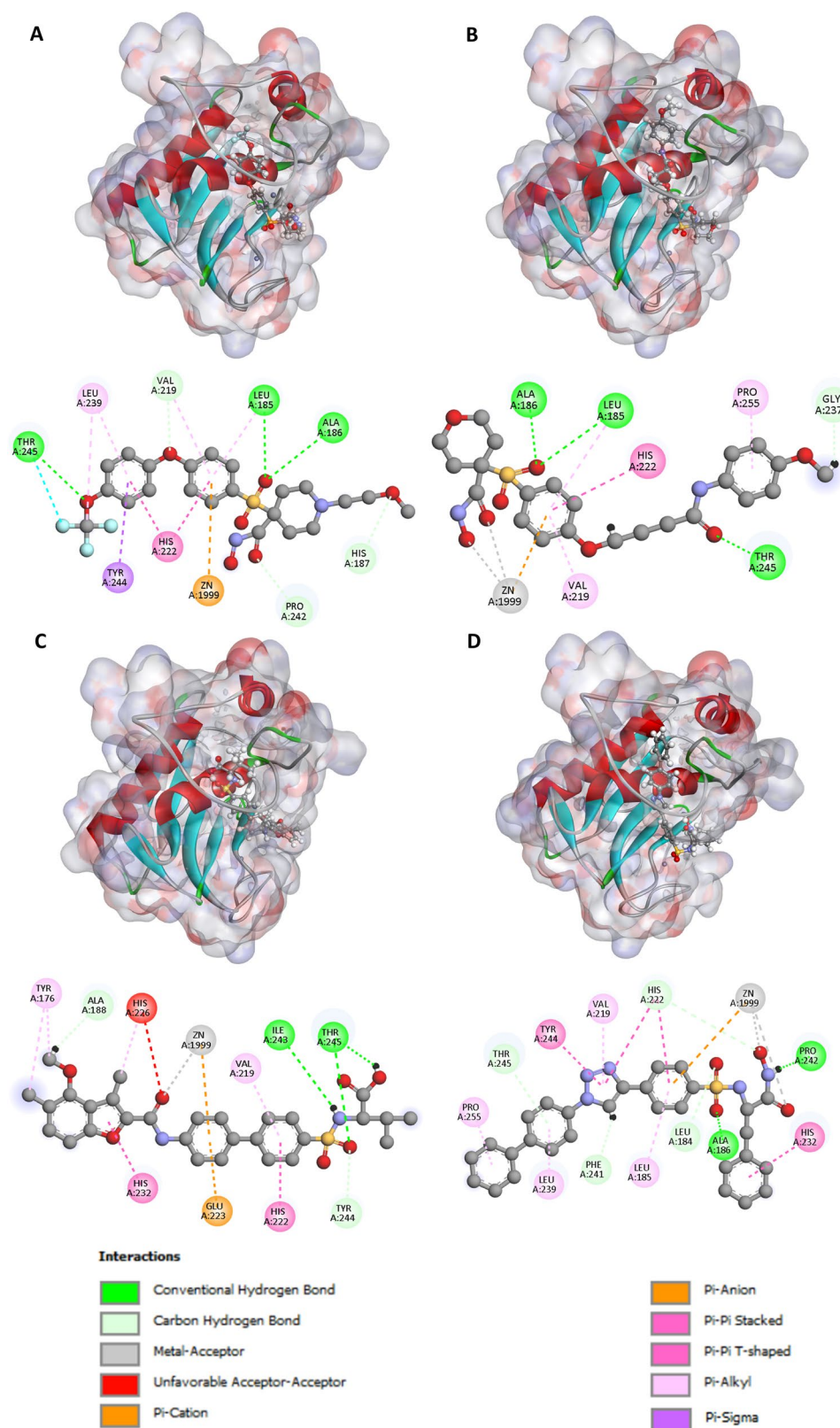
highlights key interactions where H-bond denotes conventional hydrogen bonds, and C-H bond refers to carbon-hydrogen bonding interactions

particularly after 100 ns, indicating greater conformational flexibility or less stable ligand binding. These observations suggest that ChEMBL1770157 forms the most stable complex with the protein over the simulation period.

The Rg plot illustrates the time-dependent compactness of the protein structure. All systems maintain Rg values between 1.52 and 1.65 nm, suggesting overall structural integrity throughout the simulation. ChEMBL1770157

and 3KRY exhibit the lowest and most stable Rg values (~1.54–1.58 nm), indicating that the protein maintains a compact and folded structure during the simulation. ChEMBL425020 demonstrates higher Rg fluctuations (up to ~1.67 nm), especially in the initial 100 ns, indicating temporary loosening or expansion of the protein structure. ChEMBL5182668 maintains intermediate compactness with slight deviations (Fig. 8B). Overall, the Rg trends

Fig. 7 Docking interactions of four selected MMP-13 inhibitors within the active site: **A** (3KRY), **B** (ChEMBL1770157), **C** (ChEMBL425020) and **D** (ChEMBL5182668). Top panels show 3D binding conformations; bottom panels depict 2D interaction maps highlighting hydrogen bonds, π -interactions, halogen bonding, and key interacting residues



support the RMSD and RMSF analyses, highlighting the superior structural compactness and stability of the ChEMBL1770157 complex.

The RMSF analysis reveals residue-wise fluctuations, highlighting the flexible and rigid regions of the protein during the simulation. Overall, all complexes exhibit

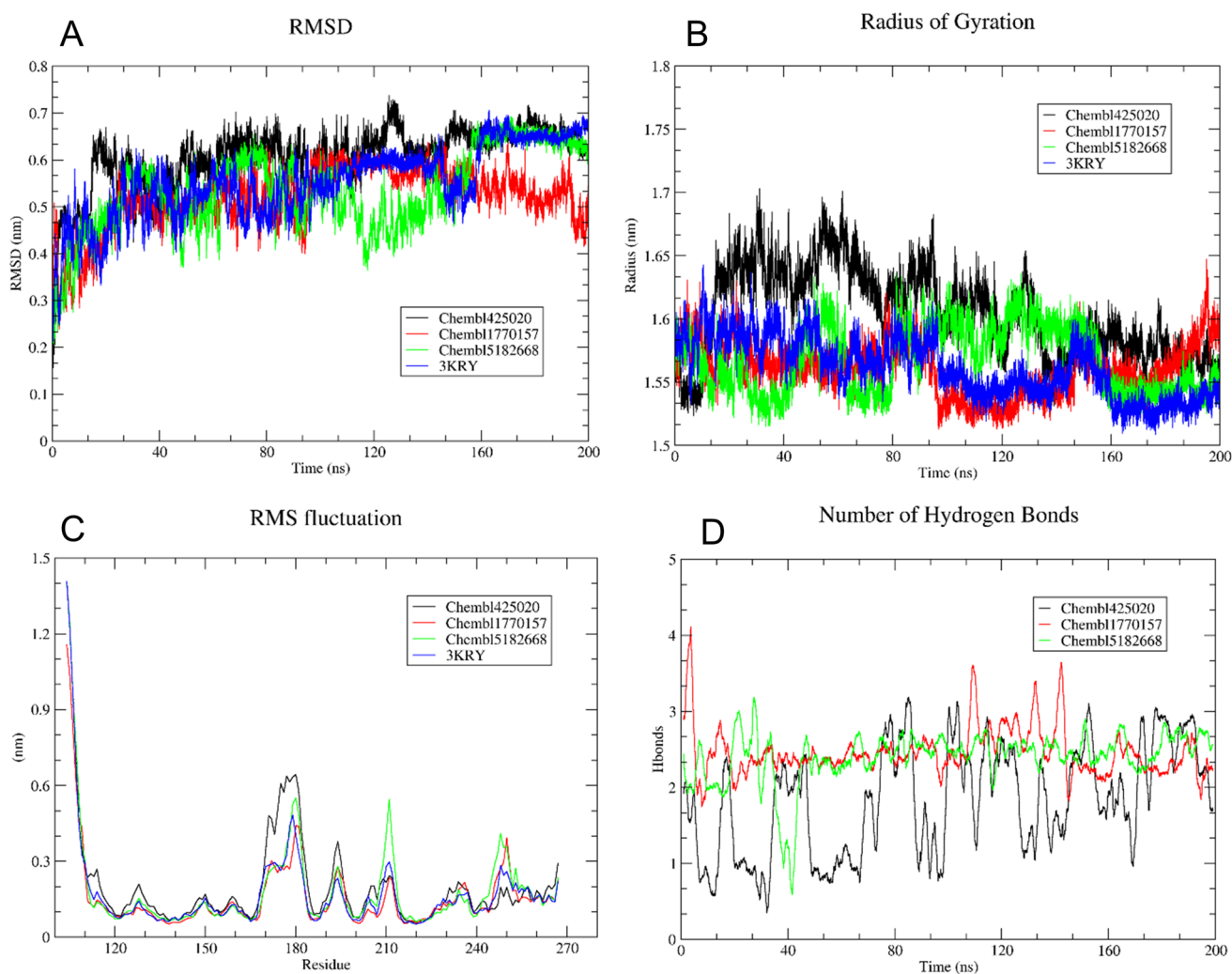


Fig. 8 Molecular dynamics (MD) simulation analysis of the MMP-13–ligand complex over a 200 ns trajectory. **A** Root Mean Square Deviation (RMSD) plot showing the structural stability of the protein–ligand complex. **B** Radius of gyration (Rg) depicting the compactness of the protein structure during the simulation. **C** Root

Mean Square Fluctuation (RMSF) plot indicating residue-level flexibility. **D** Hydrogen bond analysis revealing the number and stability of intermolecular hydrogen bonds between the ligand and MMP-13 throughout the simulation

a similar fluctuation profile with moderate flexibility (~ 0.1 – 0.6 nm) across most residues. Higher fluctuations are observed near the N-terminal region (residues ~ 115 – 125), potentially due to solvent exposure or loop regions. Binding site residues (typically in the central domain) show relatively low RMSF values (< 0.3 nm), indicating that ligand binding induces rigidity and stabilizes the local conformation. ChEMBL425020 displays slightly elevated fluctuations across certain residues compared to other complexes, aligning with its higher RMSD values (Fig. 8C). The number of hydrogen bonds formed between the ligand and the protein during the MD simulation provides insight into the interaction stability and binding affinity. For the ChEMBL1770157 and ChEMBL5182668 complexes, persistent hydrogen bonds were observed (Fig. 8D),

indicating strong and sustained interactions with the binding pocket. These ligands consistently formed 2–3 hydrogen bonds throughout the 200 ns trajectory, supporting their favorable binding profiles and conformational stabilization of the protein. In contrast, the ChEMBL425020 complex exhibited sporadic hydrogen bonding, with noticeable fluctuations and intermittent loss of interactions, suggesting a weaker or more transient binding mode.

Interestingly, no hydrogen bonds were detected between the protein and ligand (3KRY). This was due to the absence of donor and acceptor atoms in the ligand as recognized by GROMACS, possibly resulting from either its chemical nature or how its topology was defined in the force field. As a result, the standard utility (gmx hbond) could not compute hydrogen bonding for this system. Despite this,

Table 8 Binding Free Energy (kcal/mol) Calculations Using MM/PBSA

Compounds	VDWAALS	EEL	EPB	ENPOLAR	GGAS	GSOLV	TOTAL
3KRY	-39.94 ± 5.44	4.08 ± 5.12	25.32 ± 14.71	-3.9 ± 0.35	-35.87 ± 7.31	21.42 ± 14.43	-14.44 ± 11
425020	-57.47 ± 4.58	-25.43 ± 9.66	58.8 ± 11.65	-5.33 ± 0.28	-82.9 ± 11.13	53.47 ± 11.52	-29.43 ± 11.65
1770157	-53.23 ± 3.21	-23.91 ± 6.28	47.28 ± 8.15	-4.94 ± 0.19	-77.14 ± 7.06	42.34 ± 8.09	-34.8 ± 5.35
5182668	-62.95 ± 5.27	-7.56 ± 8.85	73.27 ± 13.85	-6.12 ± 0.3	-70.51 ± 10.54	67.14 ± 13.89	-3.37 ± 14.04

* calculations represent average $\pm$ Std. deviation

other parameters such as RMSD and Rg indicate that the complex maintains structural stability, hinting that non-hydrogen-bonding interactions (e.g. hydrophobic contacts, π -stacking) may play a dominant role in ligand binding in this case.

MM/PBSA analysis

The energetic underpinnings of ligand binding were investigated using MM/PBSA calculations on representative MD simulations (Table 8). This analysis yielded the contributions of van der Waals (VDWAALS), electrostatic (EEL), polar solvation (EPB), and nonpolar solvation (ENPOLAR) energies to the overall binding free energy. As depicted in Fig. 9, which represents averages over the final 200 ns MD frames, these energy terms play distinct roles in the stability of the ligand–protein complexes. Among the evaluated ligands, ChEMBL1770157 exhibited the most favorable total binding free energy (-34.8 ± 5.35 kcal/mol), with significant contributions from both van der Waals (-53.23 ± 3.21 kcal/mol) and electrostatic interactions (-23.91 ± 6.28 kcal/mol) moderately offset by solvation energies. ChEMBL425020 followed closely (-29.43 ± 11.65 kcal/mol), also showing strong nonbonded interaction terms. In contrast, ChEMBL5182668, although it displayed the strongest van der Waals contribution (-62.95 ± 5.27 kcal/mol), suffered from a large unfavorable polar solvation energy (73.27 ± 13.85 kcal/mol), resulting in a negligible total binding energy (-3.37 ± 14.04 kcal/mol). This suggests that desolvation penalties and lack of sufficient electrostatic stabilization undermine its binding efficacy. The 3KRY served as a control and showed weak binding affinity (-14.44 ± 11 kcal/mol), further supporting the contribution of ligand-induced stabilization observed in the active complexes.

Residue-level energy decomposition highlighted critical hotspots mediating ligand recognition and stabilization. For ChEMBL1770157, residues such as Glu223, Phe241, and Leu2018 contributed significantly to binding via electrostatic and hydrophobic interactions, reflecting optimal complementarity with the ligand's chemical features.

Similarly, ChEMBL425020 showed strong interactions with Glu223, Ala238, and Phe252. In contrast, the interaction profile of ChEMBL5182668 revealed destabilizing contributions from His222 and Glu223, indicating an electrostatic mismatch that could explain the poor net binding affinity. Importantly, Leu185, Glu223, Ala238, and Tyr244 emerged as recurring electrostatic hotspots across all active complexes, emphasizing their potential as key anchoring residues for ligand optimization. Collectively, these findings underscore that strong van der Waals forces must be accompanied by favorable electrostatic and solvation properties to yield high-affinity binding, and that ligand design efforts should prioritize interactions with conserved hotspots like Glu223 and Tyr244 to enhance binding potency and specificity.

By dissecting the total binding free energy into per-residue contributions, particularly the van der Waals and electrostatic components, this analysis allowed for the pinpointing of critical residues, often referred to as “hotspots” that are essential for stabilizing the protein–ligand interaction. The insights gained from combining MM/PBSA calculations with per-residue decomposition analysis were invaluable for understanding the intricate molecular mechanisms that govern MMP13–ligand interactions and for guiding the rational design of novel therapeutic agents.

Conclusions

This study provides a comprehensive computational analysis of MMP-13 inhibitors by integrating QSAR modeling, machine learning, docking, and MD simulations. Among eight regression models, LGBM, SVR, and RF showed robust predictive performance, with LGBM offering the best trade-off between accuracy and generalization. Classification models such as SVM and LGBM further demonstrated high predictive reliability in distinguishing active inhibitors, fulfilling OECD validation criteria. Docking revealed key interactions, such as hydrogen bonding and π – π stacking, with three compounds exhibiting high binding affinity. Docking revealed that all inhibitors consistently interacted

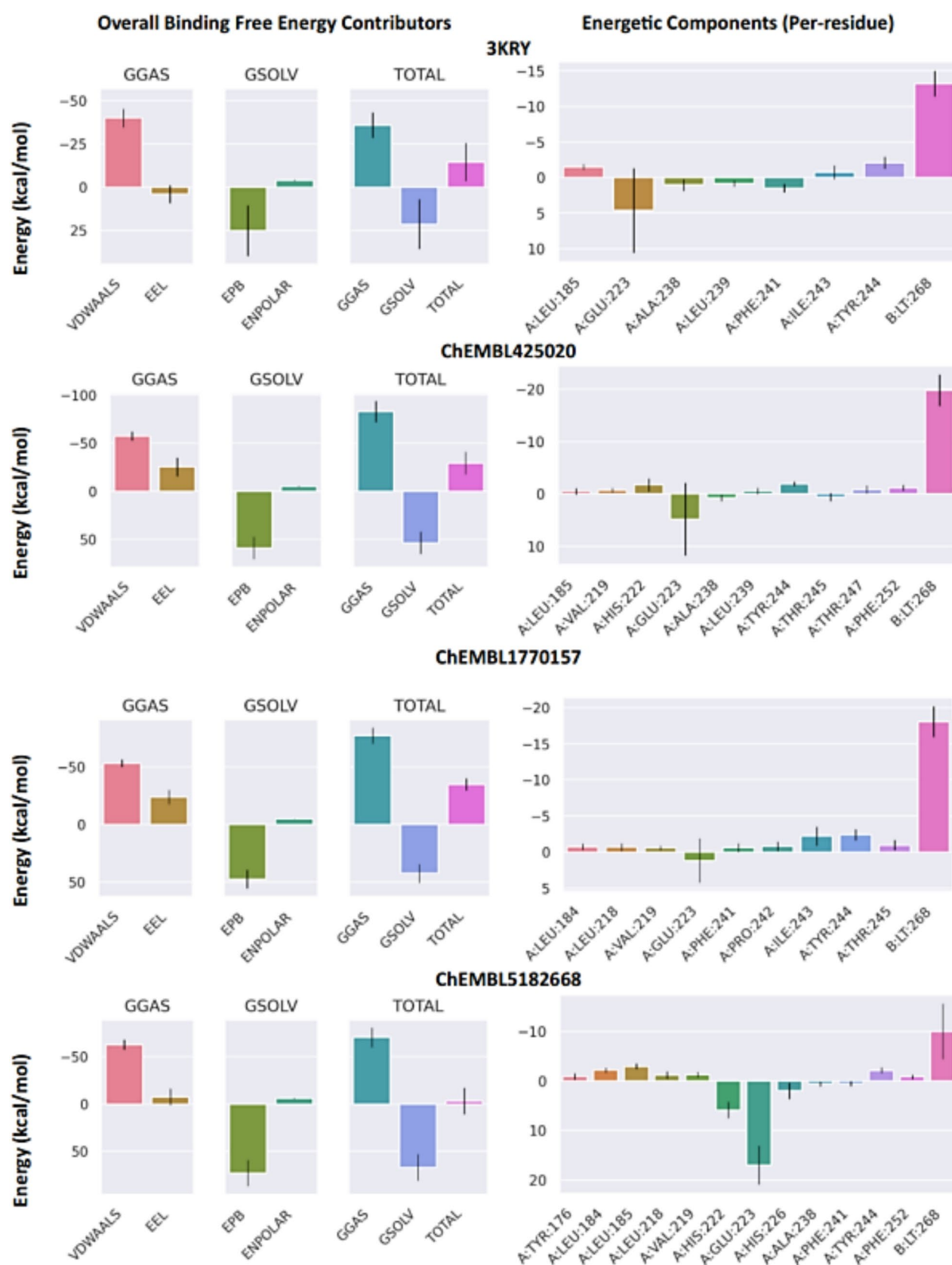


Fig. 9 MM/PBSA results (left) binding free energy contribution by various interactions, (right) binding free energy contribution by active residues and ligands

with a core set of MMP-13 active site residues. These top candidates were further validated through 200 ns MD simulations, confirming their structural stability and favorable dynamic behavior within the MMP-13 active site. These findings support structure-guided optimization for designing potent and selective MMP-13 inhibitors. Overall, the study underscores the critical role of physicochemical and structural properties in MMP-13 inhibitor design and highlights promising leads for the development of novel therapeutics targeting MMP-13-mediated inflammatory disorders. This study lays a solid foundation for the rational design of novel MMP-13 inhibitors and their potential translation into therapeutic agents.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11030-025-11358-5>.

Author contributions A.M. and S.I. led the conceptualization, methodology development, software implementation, validation, formal analysis, and investigation. S.I. was also responsible for data curation, original draft preparation, review and editing, visualization, and project administration. E.K. was involved in manuscript review and editing, and project administration for the study. S.C. and D.L. were involved in manuscript review and editing, project administration, and secured funding for the study.

Funding This research is supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00523080 and NRF-2023R1A2C2003034).

Data availability The dataset used in this study is available at <https://github.com/SI319>.

Declarations

Conflict of interest The authors declare no competing interests.

References

- Page-McCaw A, Ewald AJ, Werb Z (2007) Matrix metalloproteinases and the regulation of tissue remodelling. *Nat Rev Mol Cell Biol* 8:221–233. <https://doi.org/10.1038/nrm2125>
- Li H, Wang D, Yuan Y, Min J (2017) New insights on the MMP-13 regulatory network in the pathogenesis of early osteoarthritis. *Arthritis Res Ther* 19:248. <https://doi.org/10.1186/s13075-017-1454-2>
- Karila T, Tervahartiala T, Cohen B, Sorsa T (2022) The collagenases: are they tractable targets for preventing cartilage destruction in osteoarthritis? *Expert Opin Ther Targets* 26:93–105. <https://doi.org/10.1080/14728222.2022.2035362>
- Knäuper V, López-Otin C, Smith B et al (1996) Biochemical characterization of human collagenase-3. *J Biol Chem* 271:1544–1550. <https://doi.org/10.1074/jbc.271.3.1544>
- Ansari MY, Ahmad N, Voleti S et al (2020) Mitochondrial dysfunction in osteoarthritis and aged cartilage triggers inflammatory response and matrix degradation via ros mediated activation of JNK-MAPK/cFos-AP1 axis in chondrocytes. *Osteoarthritis Cartilage* 28:S187. <https://doi.org/10.1016/j.joca.2020.02.304>
- Hattori N, Mochizuki S, Kishi K et al (2009) MMP-13 plays a role in keratinocyte migration, angiogenesis, and contraction in mouse skin wound healing. *Am J Pathol* 175:533–546. <https://doi.org/10.2353/ajpath.2009.081080>
- Selvamurugan N, Pulumati MR, Tyson DR, Partridge NC (2000) Parathyroid hormone regulation of the rat collagenase-3 promoter by protein kinase A-dependent transactivation of core binding factor $\alpha 1$. *J Biol Chem* 275:5037–5042. <https://doi.org/10.1074/jbc.275.7.5037>
- Wang X, Manner PA, Horner A et al (2004) Regulation of MMP-13 expression by RUNX2 and FGF2 in osteoarthritic cartilage. *Osteoarthritis Cartilage* 12:963–973. <https://doi.org/10.1016/j.joca.2004.08.008>
- Hu Q, Ecker M (2021) Overview of MMP-13 as a promising target for the treatment of osteoarthritis. *Int J Mol Sci* 22:1742. <https://doi.org/10.3390/ijms22041742>
- Xu M, Zhang L, Zhao L et al (2013) Phosphorylation of osteopontin in osteoarthritis degenerative cartilage and its effect on matrix metalloproteinase 13. *Rheumatol Int* 33:1313–1319. <https://doi.org/10.1007/s00296-012-2548-4>
- Chen D, Li Y, Dai X et al (2013) 1,25-dihydroxyvitamin D3 activates MMP13 gene expression in chondrocytes through p38 MARK pathway. *Int J Biol Sci* 9:649–655. <https://doi.org/10.7150/ijbs.6726>
- Kalva S, Saranyah K, Rath S, Suganya P et al (2013) Potent inhibitors precise to S1' loop of MMP-13, a crucial target for osteoarthritis. *J Mol Graph Model* 44:297–310. <https://doi.org/10.1016/j.jmgm.2013.06.005>
- Yoshihara Y, Nakamura H, Obata K et al (2000) Matrix metalloproteinases and tissue inhibitors of metalloproteinases in synovial fluids from patients with rheumatoid arthritis or osteoarthritis. *Ann Rheum Dis* 59:455–461. <https://doi.org/10.1136/ard.59.6.455>
- Ni S, Li C, Xu N et al (2019) Follistatin-like protein 1 induction of matrix metalloproteinase 1, 3 and 13 gene expression in rheumatoid arthritis synoviocytes requires MAPK, JAK/STAT3 and NF- κ B pathways. *J Cell Physiol* 234:454–463. <https://doi.org/10.1002/jcp.26580>
- Sun G-H, Liao Y, et al (2018) Electroacupuncture intervention improves cartilage degeneration and subchondral bone osteoporosis of knee-joint possibly by adjusting OPG/RANK/RANKL signaling in ovariectomized rats. *Zhen Ci Yan Jiu*. 43:781–7. <https://doi.org/10.13702/j.1000-0607.170591>
- Wu C, Liu J, Tsai H et al (2022) Interleukin-11/gp130 upregulates MMP-13 expression and cell migration in OSCC by activating PI3K/Akt and AP-1 signaling. *J Cell Physiol* 237:4551–4562. <https://doi.org/10.1002/jcp.30902>
- Freije JM, Díez-Itza I, Balbín M et al (1994) Molecular cloning and expression of collagenase-3, a novel human matrix metalloproteinase produced by breast carcinomas. *J Biol Chem* 269:16766–16773. [https://doi.org/10.1016/S0021-9258\(19\)89457-7](https://doi.org/10.1016/S0021-9258(19)89457-7)
- Francisco V, Ait Eldjoudi D, González-Rodríguez M et al (2023) Metabolomic signature and molecular profile of normal and degenerated human intervertebral disc cells. *Spine J* 23:1549–1562. <https://doi.org/10.1016/j.spinee.2023.06.005>
- Zhu H, Chen G, Wang Y et al (2020) Dimethyl fumarate protects nucleus pulposus cells from inflammation and oxidative stress and delays the intervertebral disc degeneration. *Exp Ther Med* 20:1–1. <https://doi.org/10.3892/etm.2020.9399>
- Zhu S, Chen K, Lan Y et al (2013) Alendronate protects against articular cartilage erosion by inhibiting subchondral bone loss in ovariectomized rats. *Bone* 53:340–349. <https://doi.org/10.1016/j.bone.2012.12.044>
- Wang H, Li H, Yan Q et al (2021) Serum matrix metalloproteinase-13 as a diagnostic biomarker for cutaneous

- squamous cell carcinoma. *BMC Cancer* 21:816. <https://doi.org/10.1186/s12885-021-08566-1>
22. Baragi VM, Becher G, Bendele AM et al (2009) A new class of potent matrix metalloproteinase 13 inhibitors for potential treatment of osteoarthritis: evidence of histologic and clinical efficacy without musculoskeletal toxicity in rat models. *Arthritis Rheum* 60:2008–2018. <https://doi.org/10.1002/art.24629>
 23. Arnoczky SP, Lavagnino M, Egerbacher M et al (2007) Matrix metalloproteinase inhibitors prevent a decrease in the mechanical properties of stress-deprived tendons. *Am J Sports Med* 35:763–769. <https://doi.org/10.1177/0363546506296043>
 24. Liechti FD, Grandgirard D, Leppert D, Leib SL (2014) Matrix metalloproteinase inhibition lowers mortality and brain injury in experimental pneumococcal meningitis. *Infect Immun* 82:1710–1718. <https://doi.org/10.1128/IAI.00073-14>
 25. Elliott S, Hays E, Mayor M et al (2003) The triterpenoid CDDO inhibits expression of matrix metalloproteinase-1, matrix metalloproteinase-13 and Bcl-3 in primary human chondrocytes. *Arthritis Res Ther* 5:R285. <https://doi.org/10.1186/ar792>
 26. Moreau M, Boileau C, Martel-Pelletier J et al (2006) Licofelone reduces progression of structural changes in a canine model of osteoarthritis under curative conditions: effect on protease expression and activity. *J Rheumatol* 33:1176–1183
 27. Tropsha A (2010) Best practices for QSAR model development, validation, and exploitation. *Mol Inform* 29:476–488. <https://doi.org/10.1002/minf.201000061>
 28. Verma RP, Hansch C (2007) Matrix metalloproteinases (MMPs): chemical-biological functions and (Q)SARs. *Bioorg Med Chem*. <https://doi.org/10.1016/j.bmc.2007.01.011>
 29. Gupta SP, Kumaran S (2005) A quantitative structure-activity relationship study on some series of anthranilic acid-based matrix metalloproteinase inhibitors. *Bioorg Med Chem*. <https://doi.org/10.1016/j.bmc.2005.05.055>
 30. Huang S, Feng K, Ren Y (2019) Molecular modelling studies of quinazolinone derivatives as MMP-13 inhibitors by QSAR, molecular docking and molecular dynamics simulations techniques. *Medchemcomm* 10:101–115. <https://doi.org/10.1039/C8MD000375K>
 31. Jahan A, Sharma BK, Sharma VD (2021) Quantitative structure-activity relationship study on the MMP-13 inhibitory activity of fused pyrimidine derivatives possessing a 1,2,4-Triazol-3-yl group as a ZBG. *GSC Biol Pharmaceut Sci*. <https://doi.org/10.30574/gscbps.2021.16.1.0199>
 32. Fjodorova N, Novich M, Vrachko M et al (2008) Directions in QSAR modeling for regulatory uses in OECD member countries, EU and in Russia. *J Environ Sci Health C Environ Carcinog Ecotoxicol Rev* 26:201–236. <https://doi.org/10.1080/10590500802135578>
 33. Lipinski CA (2004) Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discov Today Technol* 1:337–341. <https://doi.org/10.1016/j.ddtec.2004.11.007>
 34. Kaneko H (2024) Evaluation and optimization methods for applicability domain methods and their hyperparameters, considering the prediction performance of machine learning models. *ACS Omega* 9:11453–11458. <https://doi.org/10.1021/acsomega.3c08036>
 35. Yap CW (2011) PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. *J Comput Chem* 32:1466–1474. <https://doi.org/10.1002/jcc.21707>
 36. Nantasenamat C (2020) Best Practices for Constructing Reproducible QSAR Models. In: *Methods in Pharmacology and Toxicology*. pp 55–75
 37. Chicco D, Jurman G (2020) The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* 21:6. <https://doi.org/10.1186/s12864-019-6413-7>
 38. Vilar S, Cozza G, Moro S (2008) Medicinal chemistry and the molecular operating environment (MOE): application of QSAR and molecular docking to drug discovery. *Curr Top Med Chem* 8:1555–1572. <https://doi.org/10.2174/156802608786786624>
 39. Huang J, MacKerell AD (2013) CHARMM36 all-atom additive protein force field: validation based on comparison to NMR data. *J Comput Chem* 34:2135–2145. <https://doi.org/10.1002/jcc.23354>
 40. Bussi G, Donadio D, Parrinello M (2007) Canonical sampling through velocity rescaling. *J Chem Phys* 126:014101. <https://doi.org/10.1063/1.2408420>
 41. Essmann U, Perera L, Berkowitz ML et al (1995) A smooth particle mesh Ewald method. *J Chem Phys* 103:8577–8593. <https://doi.org/10.1063/1.470117>
 42. Abraham MJ, Murtola T, Schulz R et al (2015) GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1:19–25. <https://doi.org/10.1016/j.softx.2015.06.001>
 43. Naveja JJ, Vogt M, Stumpfe D et al (2019) Systematic extraction of analogue series from large compound collections using a new computational compound-core relationship method. *ACS Omega* 4:1027–1032. <https://doi.org/10.1021/acsomega.8b03390>
 44. Moi D, Nocentini A, Deplano A et al (2019) Structure-activity relationship with pyrazoline-based aromatic sulfamates as carbonic anhydrase isoforms I, II, IX and XII inhibitors: synthesis and biological evaluation. *Eur J Med Chem* 182:111638. <https://doi.org/10.1016/j.ejmech.2019.111638>
 45. Gege C, Bao B, Bluhm H et al (2012) Discovery and evaluation of a non-Zn chelating, selective matrix metalloproteinase 13 (MMP-13) inhibitor for potential intra-articular treatment of osteoarthritis. *J Med Chem* 55:709–716. <https://doi.org/10.1021/jm201152u>
 46. Perrone MG, Blevé L, Santandrea E et al (2009) The tertiary amine nitrogen atom of piperazine sulfonamides as a novel determinant of potent and selective β_3 -adrenoceptor agonists. *ChemMedChem* 4:2080–2097. <https://doi.org/10.1002/cmdc.200900292>
 47. La Regina G, Puxeddu M, Nalli M et al (2020) Discovery of new 1,1'-biphenyl-4-sulfonamides as selective subnanomolar human carbonic anhydrase II inhibitors. *ACS Med Chem Lett* 11:633–637. <https://doi.org/10.1021/acsmchemlett.9b00437>
 48. Abdulkader AM, Taher M, Nik Yusoff NI, Alaama M (2021) Discovery of bis-sydnone styryl ketone as a selective cyclooxygenase-2 inhibitor. *Futur J Pharm Sci* 7:62. <https://doi.org/10.1186/s43094-021-00216-4>
 49. Meanwell NA (2013) The Influence of Bioisosteres in Drug Design: Tactical Applications to Address Developability Problems. In: *Topics in Medicinal Chemistry*. pp 283–381
 50. Chen AY, Adamek RN, Dick BL et al (2019) Targeting metalloenzymes for therapeutic intervention. *Chem Rev* 119:1323–1455. <https://doi.org/10.1021/acs.chemrev.8b00201>
 51. Sağlık BN, Osmaniye D, Çevik UA et al (2020) Synthesis, characterization and carbonic anhydrase I and II inhibitory evaluation of new sulfonamide derivatives bearing dithiocarbamate. *Eur J Med Chem* 198:112392. <https://doi.org/10.1016/j.ejmech.2020.112392>
 52. Bai R, Wei Z, Liu J et al (2012) Synthesis and biological evaluation of 4'-[(benzimidazole-1-yl)methyl]biphenyl-2-sulfonamide derivatives as dual angiotensin II/endothelin A receptor antagonists. *Bioorg Med Chem* 20:4661–4667. <https://doi.org/10.1016/j.bmc.2012.06.011>
 53. Kumaran S, Gupta S (2007) A quantitative structure-activity relationship study on some novel series of hydroxamic acid analogs acting as matrix metalloproteinase inhibitors. *Med Chem (Los Angeles)* 3:167–173. <https://doi.org/10.2174/157340607780059477>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the

author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.