Standigm
AI for new drug R&D

GT-AutoML™

Reliable QSAR, Powered by Deep Learning and Chemical Intelligence

Executive Summary

Quantitative Structure-Activity Relationship (QSAR) modeling is crucial in modern drug discovery for predicting compound activity, prioritizing candidates, and optimizing chemical series. However, developing accurate and reliable QSAR models often requires significant expertise in complex areas like data curation, feature engineering, and model validation, creating bottlenecks in the discovery pipeline. This white paper introduces **GT-AutoML™**, an **advanced automated platform** designed to overcome these challenges and accelerate the generation of high-quality, **predictive QSAR models**.

GT-AutoML™ streamlines the entire QSAR modeling process, **supporting both ligand-based and structure-based approaches** to provide versatility for various drug discovery programs. The platform automates labor-intensive steps such as ligand alignment and integrates multi-level molecular descriptors, including inputs from state-of-the-art foundation models, to build a comprehensive feature set. Leveraging the power of deep learning with its core TabPFN engine, GT-AutoML™ delivers robust, cross-validated models ready for prospective predictions, empowering researchers to rapidly generate reliable SAR insights and make data-driven decisions with minimal manual effort. Benchmarking on the widely-used **Sutherland dataset** further validates this capability, showing that GT-AutoML™ achieves an **average predictive accuracy up to 47% higher** than tested competing QSAR platforms.

Introduction

Quantitative Structure-Activity Relationship (QSAR) modeling has become an essential component in modern drug discovery, enabling researchers to predict the biological activity of chemical compounds based on their structural and physicochemical properties. In the era of big data and artificial intelligence, the demand for custom, high-performing QSAR models has grown substantially. These models support the prioritization of lead compounds, optimization of chemical libraries, and reduction of experimental costs. However, building reliable models requires expertise in structure preparation, feature engineering, model

selection, and validation—challenges that often act as barriers for non-experts. To address this, we present GT-AutoML™, an automated QSAR modeling platform designed to streamline and enhance the development of ligand-based and structure-based predictive models with minimal manual intervention, consistently delivering high accuracy as demonstrated in rigorous benchmarks.

Features of GT-AutoML™

GT-AutoML™ offers a powerful and flexible framework for QSAR modeling, integrating state-of-the-art machine learning techniques with domain-specific automation. Its key features include:

Support for Both Ligand-Based and Docking-Based Modeling

GT-AutoML™ enables the construction of models using both traditional ligand-based approaches and structure-based (docking-driven) methodologies, providing versatility across diverse screening scenarios.

Automatic Ligand Alignment

GT-AutoML™ performs automatic alignment of ligands to ensure consistent feature extraction, which is critical for accurate 3D and 4D descriptor generation. For docking-based alignment, users can select multiple PDB structures, increasing the likelihood of generating reliable poses while maintaining an acceptable computational cost.

Multi-Level Feature Integration

GT-AutoML™ incorporates a wide range of molecular features, including:

- 2D descriptors (e.g., topological and physicochemical properties)
- 3D descriptors (e.g., spatial and electrostatic features)
- 4D grid-based features (e.g., conformational ensemble representations)
- Embedding vectors from state-of-the-art chemistry foundation models, including Standigm's Foundation Model, which generates robust, generalizable representations through a modular multi-modal and multitask learning approach.
- Docking-derived features (available for docking-based models)

Deep Learning-Based Modeling with TabPFN

At its core, GT-AutoML™ leverages the TabPFN model—a high-performance deep learning approach designed for tabular data—to produce robust, generalizable QSAR models with minimal hyperparameter tuning.

Data Augmentation for Improved Reliability

The platform includes built-in data augmentation strategies to enhance model robustness and mitigate overfitting, especially when dealing with limited or noisy datasets.

Support for Ranged Labels in Regression Tasks

GT-AutoML™ is capable of handling ranged biological activity labels (e.g., "Activity of compound A is between 7–9" or "Compound B has activity < 4"), enabling flexible and realistic modeling in scenarios where exact values are not always available.

Detailed Workflow: Automated Ligand- and Docking-Based QSAR with GT-AutoML™

For automated QSAR modeling, Figure 1 illustrates the unified GT-AutoML™ workflow for end-to-end SAR model generation. The process begins with user-supplied ligands and, when available, a receptor structure. GT-AutoML™ first determines whether a high-quality receptor model exists; if so, it proceeds in Dock Mode, performing high-throughput docking, selecting the most consistent binding poses, and extracting interaction fingerprints and docking scores. If no receptor structure is present, it switches to Ligand Mode, where it generates diverse 3-D conformers for each molecule, aligns them to maximize shared pharmacophoric features, and retains the best representatives. In either path, the platform then fuses multi-level descriptors—including 2-D physicochemical properties, 3-D/4-D grid fields, docking-derived metrics (when applicable), and foundation-model embeddings—into a single feature matrix. This matrix is fed to TabPFN, a deep-learning engine optimized for tabular data, to train a cross-validated SAR model that is immediately ready for prospective prediction via GT-AutoML™. By automating pose generation, feature engineering, and model training in both structure-based and purely ligand-based settings, GT-AutoML™ compresses weeks of expert effort into a reproducible push-button workflow.

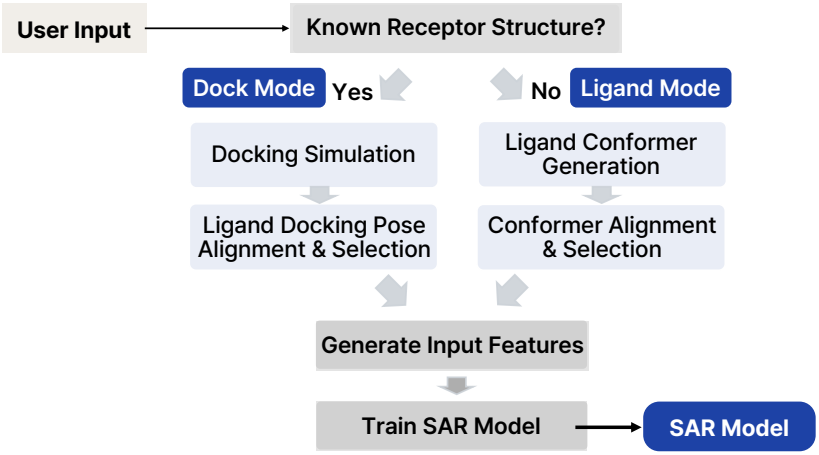

Figure 1. GT-AutoML™ decision-tree workflow for SAR model generation.

Benchmark Results

To assess the performance of GT-AutoML™, we conducted benchmark experiments using the Sutherland dataset, which comprises activity data for 8 well-characterized protein targets. For docking-based models, experimentally determined structures from the Protein Data Bank (PDB) were utilized (Table 1). These structures were selected to represent diverse protein classes, although they may not represent the optimal conformation for every ligand in the dataset. Nevertheless, they provide a consistent basis for evaluating docking-based modeling performance.

Table 1. PDB structures employed in the benchmark (relevant only for docking-based modelling). These structures were selected solely for benchmarking and may not represent the optimal choice for each target.

Target	PDB Structures
ACE	7z70A
ACHE	8aenA; 8aevA
BZR	8bgiA; 8bhgA; 8bhkA
COX2	5kirA
DHFR	6de4A
GPB	3cemM
THERM	6tmnE
THR	8aseB; 6zv8H

Figure 2 presents the Pearson R^2 values of the models across the test sets from the Sutherland dataset, with error bars indicating the standard deviation across repeated experiments. While GT-AutoML™ ligand-based models demonstrate superior performance for certain targets and docking-based models for others, the overall mean Pearson R^2 values of both GT-AutoML™ ligand- and docking-based approaches are largely comparable.

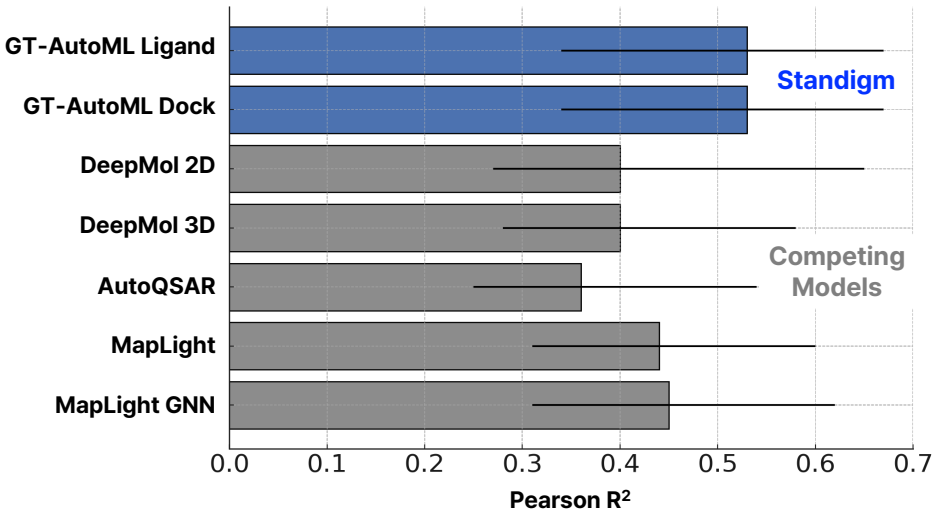

Figure 2. Predictive accuracy of GT-AutoML™ models versus established QSAR baselines on the Sutherland benchmark. Both GT-AutoML™ Ligand Mode and Dock Mode achieve comparable performance.

Figure 3 presents a comparative analysis of the average predictive performance (Pearson R^2 score) of GT-AutoML™ against several other established QSAR modeling tools on the Sutherland dataset. GT-AutoML™'s Ligand and Docking modes both achieved an average Pearson R^2 of 0.53 on this benchmark. This performance significantly surpasses that of the competing models tested, including DeepMol 2D (0.40), DeepMol 3D (0.40), AutoQSAR (0.36), MapLight (0.44), and MapLight GNN (0.45). Specifically, GT-AutoML™ demonstrates an average accuracy improvement of approximately 18% compared to the next best competitor (MapLight GNN) and up to over 47% when compared to other tools such as AutoQSAR. These results underscore the superior predictive power and reliability of the GT-AutoML™ platform for SAR model generation. Although the Sutherland dataset provides pre-aligned 3D poses for ligands, the GT-AutoML™ models were constructed using GT-AutoML™'s own automatic alignment procedure, ensuring consistency and reproducible results.

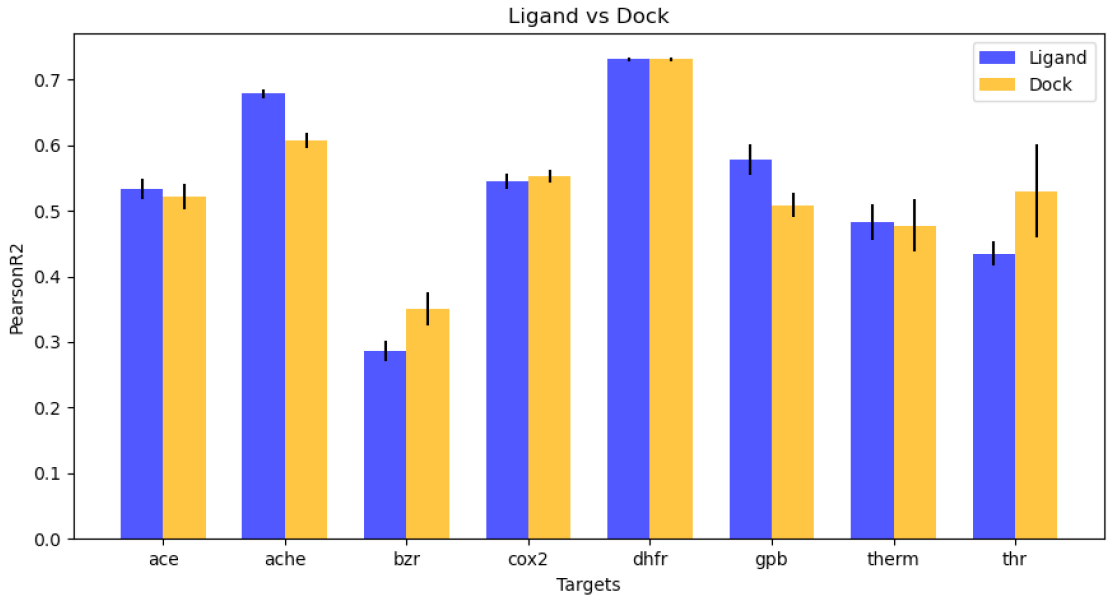

Figure 3. Per-target Pearson R^2 score for GT-AutoML™ Ligand mode (blue) and Dock mode (orange) on the Sutherland benchmark. The two modes display near-identical overall accuracy, with Dock mode outperforming on BZR and THR, and Ligand mode leading on ACE, ACHE, GPB and THERM—highlighting GT-AutoML™'s robustness whether receptor structures are available.

Conclusion

GT-AutoML™ offers a robust, automated solution for QSAR modeling, combining deep learning with domain-specific intelligence. Its flexibility across ligand- and docking-based paradigms, support for ranged labels, and high-performing TabPFN-based models make it a valuable tool for both novice and expert users. As demonstrated through benchmarking on the Sutherland dataset, GT-AutoML™ consistently achieves superior predictive performance, with average accuracy up to 47% higher than competing platforms, validating its capability to deliver reliable and impactful SAR insights. Future work will focus on boosting the representation capability of our chemistry foundation model, incorporating additional data modalities, and enhancing interpretability

For more information and to explore how GT-AutoML™ can enhance your drug discovery efforts, visit Standigm's GT-AutoML™ page and Standigm's website.

Reference

- JJ Sutherland, LA O'Brien, and DF Weaver, A Comparision of Methods for Modeling Quantitative Structure – Activity Relationships, *J Med Chem* 2004, 47, 5541-5554.

- J Correia, J Capela, and M Rocha, Deepmol: an automated machine and deep learning framework for computational chemistry, *J Cheminfo* 2024, 16, 136.

Contact Us

For more information about Standigm and GT-AutoML™ and how it can benefit you, please reach out to us:
- Email: business@standigm.com
- Website: www.standigm.com
- Latest GT-AutoML™ White Paper: https://www.standigm.com/ai-saas/gt-automl

Legal Information

About Standigm

Standigm is a Korean AI drug discovery company with over 10 years of experience, backed by SK, Pavilion Capital, and Kakao Ventures, with $70M+ in funding and partnerships with global pharma. Standigm's proprietary AI platforms—Standigm ASK™ for target identification and Standigm BEST™ for compound design—power continuous generation of commercially valuable pipelines.