JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING

pubs.acs.org/jcim Article

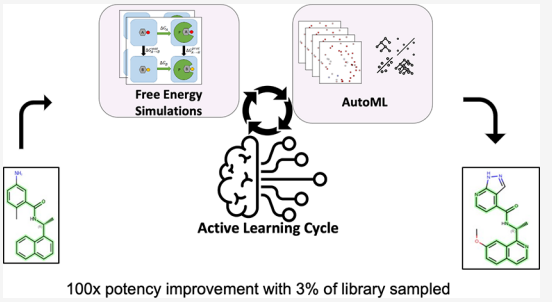
Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling

Filipp Gusev,[§] Evgeny Gutkin,[§] Maria G. Kurnikova,* and Olexandr Isayev*

Cite This: *J. Chem. Inf. Model.* 2023, 63, 583−594 Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: *In silico* identification of potent protein inhibitors commonly requires prediction of a ligand binding free energy (BFE). Thermodynamics integration (TI) based on molecular dynamics (MD) simulations is a BFE calculation method capable of acquiring accurate BFE, but it is computationally expensive and time-consuming. In this work, we have developed an efficient automated workflow for identifying compounds with the lowest BFE among thousands of congeneric ligands, which requires only hundreds of TI calculations. Automated machine learning (AutoML) orchestrated by active learning (AL) in an AL−AutoML workflow allows unbiased and efficient search for a small set of best-performing molecules. We have applied this workflow to select inhibitors of the SARS-CoV-2 papain-like protease and were able to find 133 compounds with improved binding affinity, including 16 compounds with better than 100-fold binding affinity improvement. We obtained a hit rate that outperforms that expected of traditional expert medicinal chemist-guided campaigns. Thus, we demonstrate that the combination of AL and AutoML with free energy simulations provides at least 20× speedup relative to the naïve brute force approaches.



INTRODUCTION

Hit-to-lead and lead optimization stages of drug design aim to discover lead compounds, molecules with improved binding affinity to a biological target, by altering the chemical structure of a hit molecule that has a demonstrated activity against the target. The typical lead optimization process is composed of first chemically synthesizing multiple compounds and then testing them for biological activity. This is known to be expensive and time-consuming.[1,2] Structure-based virtual screening of ultra-large molecular libraries, which aims to minimize the number of compounds chosen for laboratory synthesis and testing, has become a successful strategy in computational drug design.[3] While high hit rates have been achieved with docking ligands to target proteins, two main limitations of such approaches remain: the limited ability of docking methodologies to predict ligand binding affinity and the technological difficulty of working with libraries composed of billions of compounds.[4,5]

Unlike docking approaches, all-atom molecular dynamics (MD) simulation methods—including thermodynamics integration (TI)[6]—can predict ligand binding affinity, also termed binding free energy (BFE), with high accuracy.[7] A relative BFE (RBFE), i.e., a BFE difference between a new ligand and a lead compound, is needed in hit-to-lead and lead optimization campaigns.[8−13] However, despite recent advances in high-performance computing and the improvement of algorithms for graphical processing unit (GPU)-accelerated MD simulations, computing multiple RBFEs for a large number of compounds remains prohibitively time-consuming and technically intractable.[14]

To overcome this problem, we have developed an automated approach for a machine learning (ML)-active learning (AL) guided lead optimization process based on RBFEs computed with TI. In this approach, compounds for the TI calculations are selected with an automated ML algorithm designed to achieve two goals: (1) to efficiently enrich a set of molecules selected for TI computation with good binders and (2) to improve an ML model's prediction of the RBFEs for an entire screening library of molecules using the TI computed RBFEs. To achieve this two-fold goal, we coupled the TI RBFE calculations with an automatic machine learning (AutoML) cycle, thus eliminating a model selection bias and efficiently utilizing an information gain on each AL iteration. This AutoML−TI RBFE computational workflow allows for the identification of tight-binding ligands with a minimal number of TI RBFE calculations.

The coronavirus disease 2019 (COVID-19) pandemic caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) remains a serious threat to global public health.

Received: August 30, 2022
Published: January 4, 2023

Given that the number of unvaccinated people is still significant, as well as the rapid rate of virus mutations, efficient COVID-19 therapeutics are needed. One of the most attractive drug targets for designing COVID-19 antivirals is the SARS-CoV-2 papain-like protease (PLpro), an enzyme responsible for processing the viral polyprotein and suppressing the host immune function.[15] PLpro has 315 residues and consists of two distinct domains: a small N-terminal ubiquitin-like domain and a "thumb−palm−fingers" catalytic domain (see Figure 1).

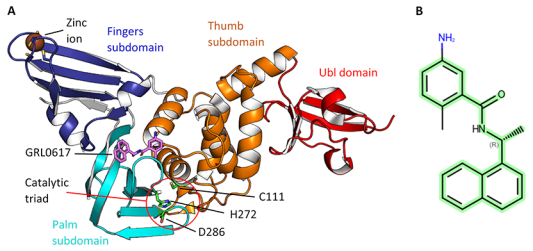


Figure 1. Structure of SARS-CoV-2 PLpro and its inhibitor. (A) Structure of SARS-CoV-2 PLpro in complex with GRL0617 (PDB ID: 7JIR). The inhibitor and residues of the catalytic triad are shown as pink and green sticks correspondingly. (B) GRL0617, with the common scaffold (*N*-[(1*R*)-1-arylethyl]arenecarboxamide) highlighted in green.

The fingers subdomain includes a zinc-binding site formed by four cysteine residues. The protein active site is formed by a canonical cysteine protease catalytic triad which includes Cys111, His271, and Asp286 residues located at an interface between the thumb and the palm subdomains.

Several *N*-[(1*R*)-1-naphthalen-1-ylethyl]benzamide derivatives were demonstrated to be effective at halting SARS-CoV-2 PLpro activity as well as SARS-CoV-2 replication in cells.[16,17] In particular, the most potent inhibitor, GRL0617, demonstrated a half-maximal inhibitory concentration (IC_{50}) of 2.3 μM,[16] and a high-resolution structure was obtained (Figure 1). In a more recent study, this inhibitor demonstrated an IC_{50} of 1.61 μM and a dissociation constant of 2.70 μM.[17] High-resolution structures of SARS-CoV-2 PLpro complexes with three inhibitors with the same scaffold, including the inhibitor GRL0617, were resolved with the resolution of 2.1−2.9 Å, revealing identical binding modes.[16] This indicates that this scaffold is important for ligand binding to PLpro and suggests that more potent PLpro inhibitors may be found among compounds with this scaffold. Recently, several novel PLpro inhibitors based on the similar scaffold were proposed with limited success using expert-driven lead optimization approaches.[17] In this work, we virtually screened a library of 1.3 billion commercially available compounds, selected a focused library of 10,000 derivatives of *N*-[(1*R*)-1-arylethyl]-arenecarboxamide, and finally identified 16 potent binders with more than 100-fold improvement in predicted binding affinity.

METHODS

Database Screening and Molecular Preparation. *Dataset.* Catalogs of three chemical vendors, WuXi [68.98 M], Mcule [2.65 M], and Enamine [1211.72 M], were combined into a single dataset, totaling approximately 1.3 billion purchasable molecules. The processing of the focused library was organized into 6 steps (Figure 2): (1) virtual screening of molecules based on the SMARTS pattern (see Figure 2A) with a Bemis−Murcko scaffold[18] of a reference ligand modified to allow substitutions of carbons to nitrogens. (2) Enumerating protonation, tautomeric, and stereoisomeric states using the OpenEye QUACPAC and OEFlipper toolkit,[19] allowing up to 8 chiral centers: only neutral molecules were retained for further calculations. (3) Filtering using SMARTS pattern (see Figure 2B): only molecules with the same configuration of the chiral center as the reference ligand (R-stereoisomer) were retained. (4) Generating 3D conformers for each of the remaining 9998 molecules using the OpenEye Omega toolkit.[20] (5) Molecular docking of generated conformers and pose filtering (see the next section for details).

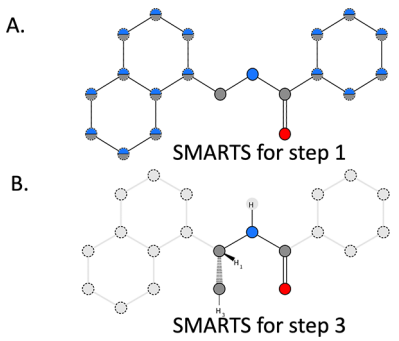


Figure 2. SMARTS filters used for small molecule dataset mining. (A) Visualization of a step 1 SMARTS pattern, used for the non-stereospecific search of *N*-[(1*R*)-1-arylethyl]arenecarboxamide derivatives. (B) Visualization of a step 3 SMARTS pattern, used for the stereospecific search of *N*-[(1*R*)-1-arylethyl]arenecarboxamide derivatives. Aromatic atoms are represented as dashed circles and aliphatic atoms as solid circles. Dark gray color represents carbon atoms, red color represents oxygen atoms, blue color represents nitrogen atoms, and light gray color represents any atom.

Molecular Docking. The representative structure of PLpro in complex with the reference ligand obtained by MD simulations (see details in the Molecular Dynamics Simulations Section) was prepared for template docking using the OpenEye Make Receptor program (version 4.0.0.0).[21] A rectangular box with edge lengths 17 × 25 × 18 Å centered on the reference ligand was specified. The outer contour of 3300 Å^3 was set and the inner contour was disabled. Three water molecules located in the binding site between the ligand and protein residues were retained, and all other water molecules were removed. The reference ligand was set as a template and no constraints were added. 3D conformers were generated from SMILES using OpenEye OMEGA (version 4.1.0.0).[20] A maximum number of conformers for a single molecule of 2000 and a minimum root mean square deviation (RMSD) of 0.2 Å were used. Template docking was performed using OpenEye HYBRID (version 4.0.0.0).[22] For each molecule, the 100 best poses were stored in the output data. All other parameters were set by default.

Pairwise atom mappings and alignment (1-to-1 atom correspondence) between a reference ligand and test ligand poses after docking, needed for initiation of MD simulations, were prepared using LS-align.[23] A molecular pose was chosen for an MD simulation if it had the lowest Hybrid Docking score and satisfied the following criteria (see Figure S1): (1) $RMSD_{core} \leq 1.3$ Å (an $RMSD_{core}$ is an RMSD between atoms of the SMARTS pattern of a test molecule and a reference ligand); (2) the length of alignment was 20 or more heavy atoms, and (3) CG3 Clash ≤0.5, where CG3 Clash is the

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

penalty term of the Chemgauss3 scoring function, which accounts for clashes between ligand and heavy protein atoms.[22] This led to a focused library of 8175 molecules.

Molecular Dynamics Simulations. *Protein System Preparation and Simulation.* The crystal structure of PLpro in complex with the reference ligand was extracted from the Protein Data Bank (PDB ID: 7JIR).[24] The protein, ligand, and zinc ion bound to protein and water molecules in the binding site of the ligand were retained, and all other molecules present in the crystal structure were removed. The input coordinates, topology, and parameters for conventional MD simulations were obtained using Ambertools 18.[25] The protein, zinc ion, and water were parameterized using FF14SB,[26] ZAFF,[27] and TIP3P[28] models, respectively. Ligand atom parameters were obtained using GAFF (version 2.11),[29] and ligand atomic charges were derived using the AM1-BCC method.[30] GPU-accelerated MD simulations were performed using the pmemd.cuda module of AMBER 18.[25] The simulation protocol included the following steps: (1) 2000 steps of minimization with the steepest descent method; (2) 100 ps of heating from 1 to 298 K in the NVT ensemble; (3) 300 ps of density equilibration in the NPT ensemble; (4) 50 ns of production simulation in NVT. Harmonic RMSD restraints were imposed on heavy atoms of the protein, ligand, and three water molecules located in the binding site during minimization and heating and were gradually removed during density equilibration. No restraints were used during production simulations.

Alchemical Relative Binding Free Energy Calculations. In this work, we computed RBFE for selected ligands with respect to the reference ligand using the alchemical thermodynamic cycle (Figure 3) reported elsewhere.[31] The RBFE is defined as a difference of the standard binding free energies (see eqs S2 and S3) of a target ligand B and a reference ligand A:

$$\Delta\Delta G_{A\to B} = \Delta G_B - \Delta G_A, \quad (1)$$

In practice, the RBFE is calculated as a difference of free energies of transforming a reference ligand into a target ligand in protein $\Delta G^{\mathrm{prot}}_{A\to B}$ and in solvent (water) $\Delta G^{\mathrm{wat}}_{A\to B}$:

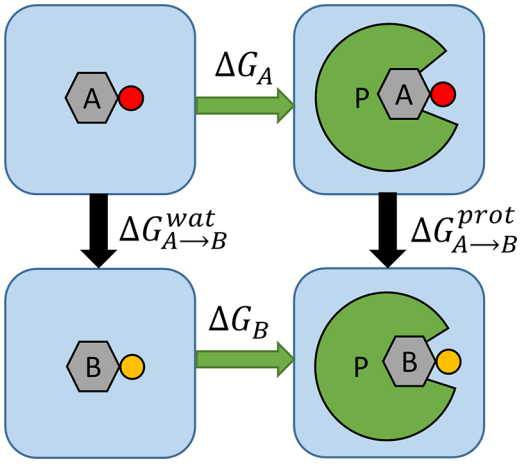


Figure 3. Thermodynamic cycle for alchemical RBFE calculations. The common substructure of ligands is shown by a gray hexagon, and the unique atoms are shown by red and yellow circles. The protein is shown by green. ΔG_A and ΔG_B correspond to standard binding free energies for the ligands A and B. $\Delta G^{\mathrm{wat}}_{A\to B}$ and $\Delta G^{\mathrm{prot}}_{A\to B}$ corresponds to the free energy of mutating ligand A to B in water and protein.

$$\Delta\Delta G_{A\to B} = \Delta G^{\mathrm{prot}}_{A\to B} - \Delta G^{\mathrm{wat}}_{A\to B}, \quad (2)$$

In this work, the free energies $\Delta G^{\mathrm{prot}}_{A\to B}$ and $\Delta G^{\mathrm{wat}}_{A\to B}$ were calculated by TI (see eq S5) using GPU-accelerated TI implementation of AMBER 18.[25] The PLpro inhibitor GRL0617 (see Figure 1) was used as a reference ligand for all TI calculations.

Ligand Preparation and Parameterization. The FESetup tool (version 1.2.1)[32] was used to set up the systems for all TI simulations. The input data for the setup procedure were ligand poses obtained by docking, the representative structure of PLpro in complex with the reference ligand obtained by MD, and the atom mappings generated by LS-align.[23] The output data were AMBER input coordinates and topologies for the solvated ligand system and the protein−ligand complex system. All atom parameters were obtained using the same force fields and charge derivation method as described in the previous section.

The first step of the setup procedure was ligand parameterization, for which the docked pose was used for the target ligand and the reference ligand was extracted from the representative structure. The target ligand was then aligned to the reference ligand using LS-align atom mappings, which outputted the pair of ligands in a vacuum. For the preparation of the solvated ligand systems, the pair of ligands was solvated in a rectangular water box with a minimum distance between the edges of the box and the ligand of 12 Å. For the preparation of the protein−ligand complex system, the pair of ligands was placed in the protein binding site of the representative structure using the coordinates of the reference ligand. Since the Cartesian coordinates of the reference ligand do not change during extraction from the representative structure, inserting the pair of ligands into the representative structure provided the same binding mode as the representative structure had initially. The obtained AMBER input coordinates and topologies were visually checked using PyMOL (version 1.8.4.0).[33]

TI Simulations. For both solvated ligand and protein−ligand complex systems, TI simulations were performed at 9 lambdas of a Gaussian quadrature (0.01592, 0.08198, 0.19331, 0.33787, 0.5, 0.66213, 0.80669, 0.91802, 0.98408). For each lambda, the system was minimized and then equilibrated using the same protocol as described in the previous section. A 4.5 ns production simulation in the NVT ensemble was then performed. The unique atoms of the target ligand and the reference ligand were modeled using soft-core potentials for both electrostatic and Van der Waals interactions. The long-range interaction cut-off, temperature, and pressure settings were the same as described in the previous section.

The orientation of ligands with respect to the protein was restrained using the virtual bond approach[34] in all TI simulations for protein−ligand complex systems. The carbon atoms of the naphthalene ring of the reference ligand and the Cα atoms of the Palm subdomain residues were used to set restraining potentials. Force constants of 5 kcal/(mol Å^2) and 5 kcal/(mol rad^2) were used for distance and angle restraints, respectively.

Average gradients were calculated from the last 4 ns of the production simulations using the alchemlyb python library.[35] The free energies for both mutations in water and in complex with the protein were obtained by the Gaussian quadrature rule. RBFE final values were obtained according to eq 2. Errors in RBFEs were estimated using the bootstrap method[13]

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

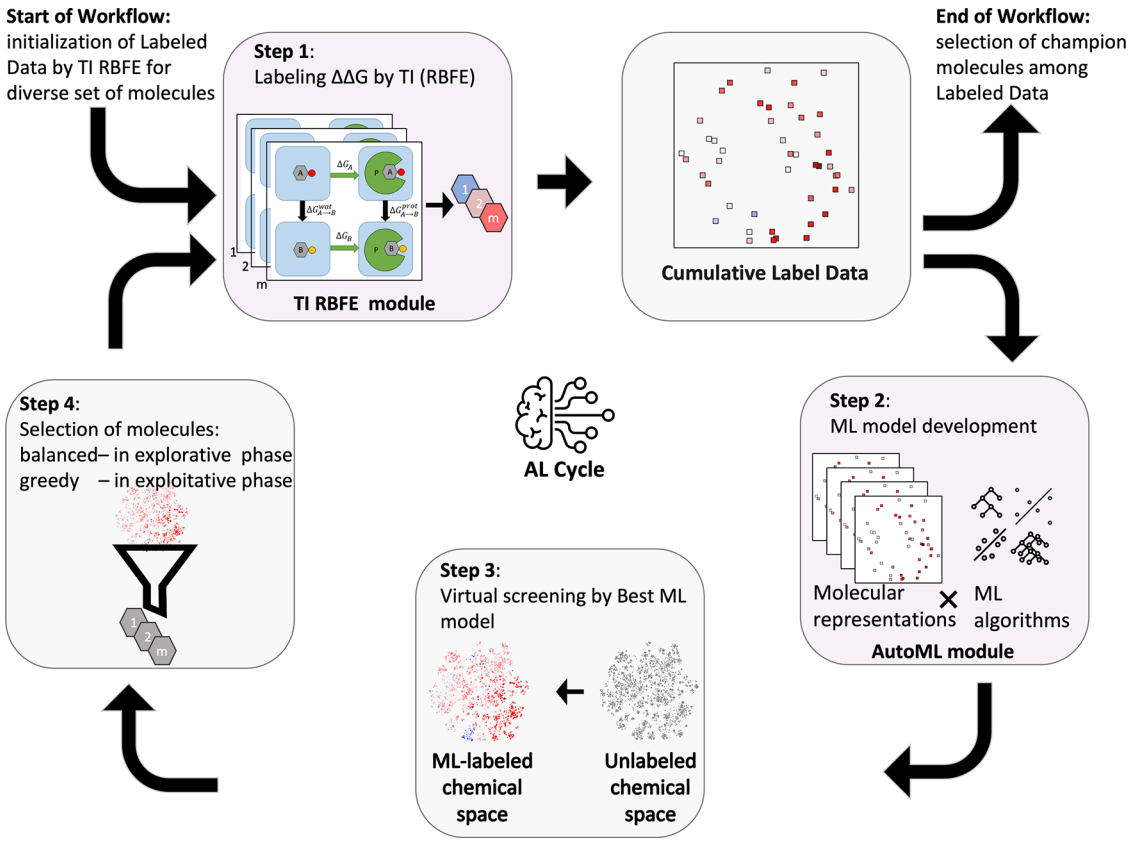


Figure 4. General scheme of the automated computational workflow organized in an active learning cycle. The workflow includes two main modules: AutoML and TI RBFE and four principal steps. Molecules with computed $\Delta\Delta G$ are depicted as colored hexagons in step 1. The labeled chemical space is shown as 2D t-SNE plots. All colors are consistent with the color scheme in Figure 7 according to their $\Delta\Delta G$ values in (blue, white, red).

implemented in the alchemlyb python library.[35] For each lambda window, the standard error in $\langle \frac{dV}{d\lambda} \rangle$ (see SI methods) was obtained by bootstrapping and then the standard error in $\Delta\Delta G$ was obtained according to the quadrature formula:

$$\sigma_{\Delta\Delta G} = \sqrt{\sum_i w_i^2 \sigma_{\langle \frac{dV}{d\lambda} \rangle_{\lambda_i}}^2}, \tag{3}$$

where λ_i is a lambda window and w_i is the weight for this window according to the Gaussian quadrature.

Converting RBFE to ΔpK_d. For a reference ligand A and a target ligand B, ΔpK_d is defined as

$$\Delta pK_d = pK_{d,B} - pK_{d,A}, \tag{4}$$

where $pK_{d,A}$ and $pK_{d,B}$ are negative decimal logarithms of dissociation constants for the reference and the target ligands, respectively. The RBFEs $\Delta\Delta G$ were converted to ΔpK_d using the following equation:

$$\Delta pK_d = -0.434 \frac{\Delta\Delta G}{RT} \tag{5}$$

where R is the gas constant, T is the temperature (298.15 K), and the coefficient of 0.434 comes from converting a natural logarithm to a decimal logarithm.

Benchmarking TI Protocol. MD input preparation and TI simulations for the benchmark set of ligands were performed following the procedures described above. Experimental absolute binding free energies (ABFE) ΔG_{exp} were obtained from the dissociation constants K_d reported in Shen et al.[17] according to the following equation:

$$\Delta G_{exp} = RT \ln \frac{K_d}{C^0}, \tag{6}$$

where C^0 is the standard concentration (1 mol/L). For each ligand, the experimental RBFE was obtained according to the following equation:

$$\Delta\Delta G_{exp} = \Delta G_{exp} - \Delta G_{exp}^{ref}, \tag{7}$$

where ΔG_{exp}^{ref} is the experimental ABFE of the reference ligand (−7.60 kcal/mol). RBFEs $\Delta\Delta G_{TI}$, computed by TI, were transformed to the ABFE ΔG_{TI} according to the following equation:

$$\Delta G_{TI} = \Delta\Delta G_{TI} + \Delta G_{exp}^{ref}. \tag{8}$$

Machine Learning Model Development and Feature Engineering. *Feature Engineering and Molecular Representation.* For the focused library of 8175 molecules (and corresponding poses), five different molecular representations were constructed: RDKit fingerprints (path fingerprints with path length 7 and binary vector length 2048) using RDKit, Morgan fingerprints (Extended-Connectivity Fingerprints with radius 3 and binary vector length 2048) using RDKit, 3D molecular fingerprints E3FP with default parameters,[36] protein−ligand extended connectivity (PLEC) fingerprints with default parameters,[37] and a combination of E3FP and PLEC fingerprints constructed as a concatenation of binary

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

vectors. For the 5 molecular representations above, three different dimensionality reduction methods (PCA, MDS, TSNE) with four reduced target dimension sizes (2, 10, 100, 200) were used. Thus, in total, 42 molecular representations were constructed (see Table S1).

AutoML. With the set of 42 possible molecular representations as an input (treated as hyperparameter), we designed the automated machine learning (AutoML) workflow based on the following classes of ML algorithms (Scikit-learn[38] implementation): Random Forest (RF), Multi-Layer Perceptron (MLP), Linear Regression (LR), KNeighbors Regression (KNR), SupportVector Regression (SVR), Gaussian Process Regression (GPR), and GaussianProcess Regression with Tanimoto Kernel (GPT). The model selection and hyperparameter search were done from scratch on each active learning cycle (see the Active Learning section) and were organized based on nested 5-fold cross-validation with mean absolute error (MAE) as the model selection criteria.

Active Learning. The active learning process was organized iteratively and consisted of eight active learning cycles (see Table S2). AL was initialized on active learning cycle 0 (AL Cycle 0) using a diverse batch of 45 representative molecules from the focused library selected by Butina-Taylor diverse clustering[39] (spherical exclusion clustering) with RDKit fingerprints (path fingerprints with path length 7 and binary vector length 2048) as a molecular representation.

AL cycles 1−5 utilized a balanced selection of molecules (for details, see Table S2, "Selection parameters"). The goal at this stage was to balance the ML model between the information gain about the chemical space and the need for selecting molecules with improved potency. In other words, those cycles were used to diversify the set of molecules in order to iteratively improve model generalizability yet keep a preference for molecules with expected negative $\Delta\Delta G$. We employed clustering to achieve this goal. The top 200 molecules with the most negative predicted $\Delta\Delta G$ were selected and then clustered into 30 clusters. From every cluster, a representative molecule with the lowest predicted $\Delta\Delta G$ was selected for RBFE calculations.

RESULTS AND DISCUSSION

Approach. The workflow developed in this work uses as a source of training data for the ML models computationally intensive MD-based thermodynamic integration calculations of the RBFEs ($\Delta\Delta G$) of a focused library of molecules with a target protein SARS-CoV-2 PLpro. To develop a focused library of compounds, we initially screened 1.3 billion commercially available molecules from three reputable compound vendors: Enamine, WuXi, and Mcule. The resulting library of approximately 10,000 *N*-[(1*R*)-1-arylethyl]-arenecarboxamide derivatives was further narrowed down to 8175 compounds, which passed structural molecular docking quality controls. These compounds were used to prepare bound poses by docking each molecule to the target binding site (see Methods, Molecular Docking subsection).

To find the best PLpro binders, we utilized active learning (AL) approach. The AL was organized as an iterative cycle (see Figure 4): (i) starting with a seed set of molecules, we performed TI RBFE calculations to train the initial ML model, (ii) we then selected molecules for the next round of the TI RBFE calculations using the current ML model, and (iii) we computed additional TI RBFEs for the molecules selected in

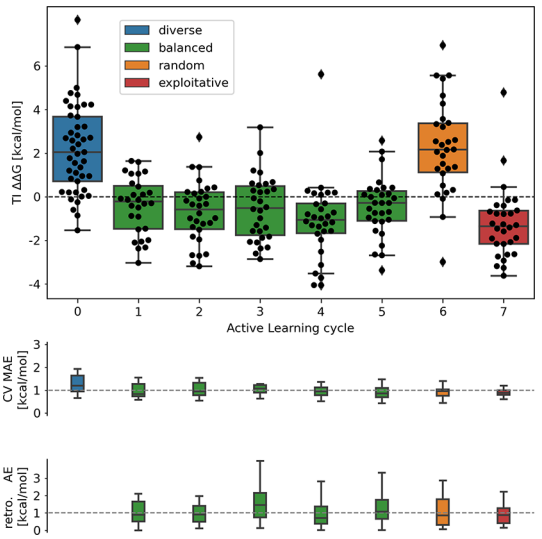


Figure 5. TI RBFE results in AL cycles (top); MAE for am ML model measured as 10-fold CV (middle); retrospective AE (bottom).

(ii) and re-trained the ML model with an updated TI RBFE dataset. The cycle was repeated until convergence.

Active Learning (AL) Cycle. The main goal of active learning is to infer an accurate ML model from a set of training data smaller than a randomly selected data set needed to achieve the same model accuracy.[40] Here, the AL workflow was organized as a black box optimization of $\Delta\Delta G$ obtained by the TI RBFE calculations for a subset of molecules from a focused library of molecules. The AL cycles were performed in two regimes: explorative and exploitative. These were distinguished by the data selection style: the explorative regime used a balanced selection, while the exploitative regime used a greedy selection (see Methods, Active Learning subsection). The explorative regime was used until the ML model reached 1 kcal/mol convergence in retrospective absolute error for two consecutive AL cycles (see Figure 5) followed by the exploitative regime used to select molecules with the lowest $\Delta\Delta G$. Both explorative and exploitative regimes were organized in an AL cycle of four steps (Figure 4): (1) train a proxy AutoML-model on acquired labeled data for a given objective(s); (2) use this model to screen the chemical space, (3) select the optimal set of candidate molecules for the TI RBFE calculations, (4) perform the TI RBFE calculations for selected molecules and use these obtained $\Delta\Delta G$ data to update the AutoML-model. The AL cycle includes two major computational modules (Figure 4): first, an AutoML module responsible for ML model development based on the labeled data provided by the second computational module, a TI RBFE module responsible for the TI computation of relative binding free energies of selected compounds in complex with the PLpro protein. The AL cycle shown in Figure 4 is initialized with a small but diverse set of molecules. Their $\Delta\Delta G$ values are computed with the TI RBFE module in step 1. These $\Delta\Delta G$s are added to the pool of labeled data, which are used by an AutoML module to train a predictive model. Specifically, labeled data are input−output pairs (X,y), where the output label y represents a correct answer to a question associated with an input X. In this work, X describes a ligand molecule, and y is a $\Delta\Delta G$ value obtained in a TI RBFE calculation (see Methods for details). An ML model trained in step 2 is used in step 3 to virtually screen the chemical space

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

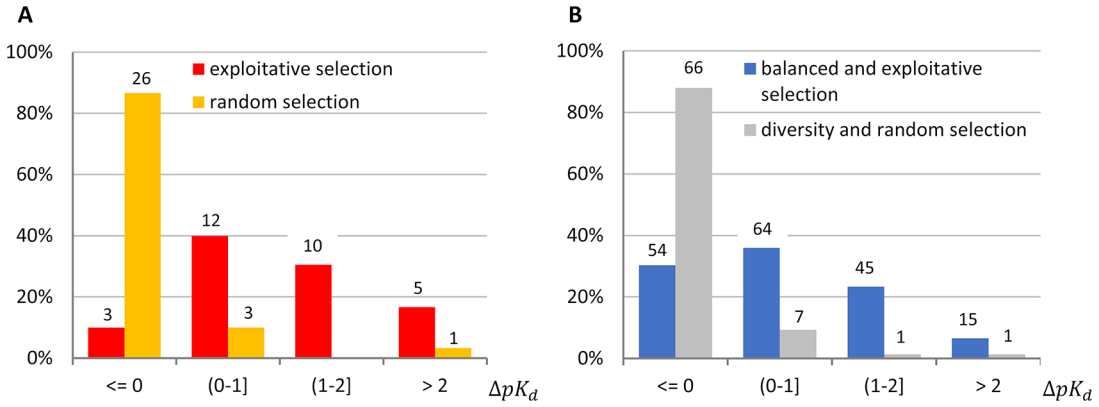


Figure 6. Normalized histograms representing the distribution of ΔpK_d for ligands selected for RBFE calculations. (A) Distribution of ΔpK_d for ligands belonged to AL explorative selection (red) and random selection (orange). (B) Distribution of ΔpK_d for ligands belonged to AL explorative and exploitative selections (blue) and diversity or random selection (gray). Definition of ΔpK_d is presented in the text. Numbers of ligands with ΔpK_d lying within the corresponding intervals are shown above bars.

(gray dots represents the chemical space) to obtain ML-model predicted $\Delta\Delta G$ values (color-coded here as in Figure 4). A new set of molecules is selected in step 4 to be submitted for the TI RBFE calculations, thus completing the cycle. The selection criteria used in step 4 are adjusted according to the specific goals: balanced for AL cycles in the explorative phase or greedy for the exploitative phase (see the Methods section for details).

Automated Machine Learning (AutoML) Module. Building an ML model with an *a priori* chosen ML method (e.g., a neural network, a Random Forest, or a Gaussian process) and a molecular representation (e.g., a path fingerprint, or a ligand−protein interaction fingerprint) may lead to a substantial model and sample selection bias. Multiple studies have shown that this bias may result in substantial modeling artifacts.[41−45] In contrast, an AutoML aims to make decisions for ML model selection, data representation, and hyperparameters in a data-driven, objective, and automated way.[46−49] The combination of AutoML and AL approaches (AutoML-AL) allows for a fast, systematic, and unbiased exploration of the chemical space in the first regime and the selection of champion candidate molecules in the second, exploitative, regime. We implemented AutoML as a set of well-performing ML algorithms available in the scikit-learn package,[38,50] multiple molecular features, feature engineering (finding optimal molecular representations), and on-the-fly selection of the best combination of an ML algorithm and molecular representation. (see Methods, Feature Engineering and Molecular Representation subsection for details).

MD-Based Thermodynamic Integration for RBFE (TI RBFE). The automated protocol for the multiple RBFE calculations implemented in this work requires minimal user interaction. Our protocol accepts a set of docked ligands (see the Methods section for details) as an input and provides calculated RBFEs for all ligands as an output. Compound GRL0617 (Figure 1B) was used as a common reference ligand. An automatic TI workflow was designed in three connected parts: (1) generation of the MD input files (including molecular topologies, initial coordinates of the atoms, and restraints), (2) set up and submission of the parallelized GPU-accelerated MD simulations using the TI implementation of the AMBER 18 package,[25] and (3) collection and processing of the output data. The details of the protocol are described in the Methods section.

Results. An AutoML−AL approach (Figure 4) was utilized to perform eight AL cycles. Figure 5 shows all TI-obtained $\Delta\Delta G$s over all AL cycles. AL cycle 0 was initialized with a diverse selection of molecules to sample the chemical space of the focused library as broadly as possible (see Methods for details). TI RBFEs were computed for this initial set of molecules and supplied to the AutoML module for initial ML-model training. For the next five AL cycles (AL cycles 1−5), we used a balanced selection of molecules from the full focused library (see Methods, Active Learning subsection for details). The goal of these five AL cycles was to gain information about the chemical space of the focused library rather than to select molecules with the lowest $\Delta\Delta G$s.

With the progression of AL cycles, the performance of the ML model improved. The cross-validated mean absolute error (MAE; see Figure 5, middle) reached 1 kcal/mol, which is comparable to the accuracy of the RBFE calculations reported elsewhere.[8−11] To verify model convergence, we performed the sixth AL cycle with a random selection of molecules (see Methods). The random selection of molecules also served to overcome the possible problem of AL being trapped in a local minimum of the chemical space.

We monitored two criteria between two subsequent cycles (AL cycles 1−6). The first criterion, the difference between the mean $\Delta\Delta G$s and the second is retrospective absolute error (AE). The difference between the mean $\Delta\Delta G$s is staying up to ca. 2 kcal/mol. The retrospective AE remains nearly constant (Figure 5, bottom) between the last balanced cycle (AL cycle 5) and the random cycle (AL cycle 6). This suggests that the AutoML−AL process converged to a desired chemical accuracy for the entire focused library. Subsequently, for the AL cycle 7, we performed an exploitative (greedy) selection of the molecules with the lowest ML predicted $\Delta\Delta G$s. The resulting AL cycle 7 had a mean TI $\Delta\Delta G$ of −1.7 kcal/mol as opposed to 2 kcal/mol in AL cycle 6 (Figure 5, top). This difference is statistically significant with the p-value = 1.3×10^{-8} according to the Mann−Whitney U test.[51]

The efficiency of the AL workflow can be further demonstrated by comparing the distributions of predicted binding affinity of the ligands selected by the ML models (including both the explorative and the exploitative sets) to those of the ligands selected by diversity or randomly (see Figure 6). For convenience, all ligand $\Delta\Delta G$s were converted to ΔpK_d (see Methods for details). Please note that $\Delta pK_d = 1$ and

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

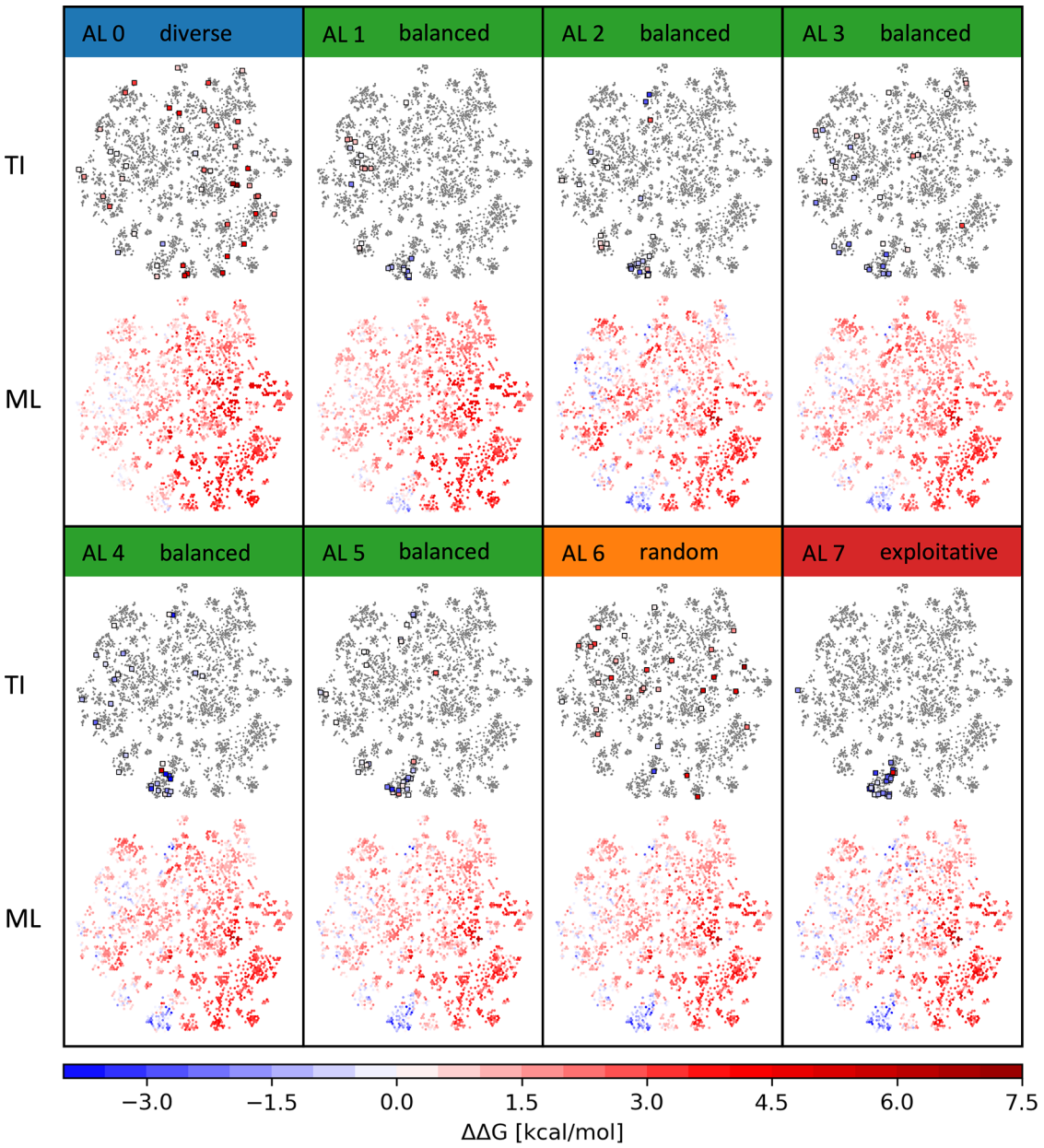


Figure 7. TI RBFE results and ML model evolution over the active learning cycles. Each panel (labeled by the AL cycle number and a corresponding selection style) shows two 2D-labeled t-SNE representations of the focused library: (top) molecules selected in a respective AL cycle for the TI RBFE calculation are colored by the TI computed ΔΔ*G*, and the rest is shown in gray; (bottom) focused library is colored by ML predicted ΔΔ*G*. The color bar for the ΔΔ*G* values is shown at the bottom.

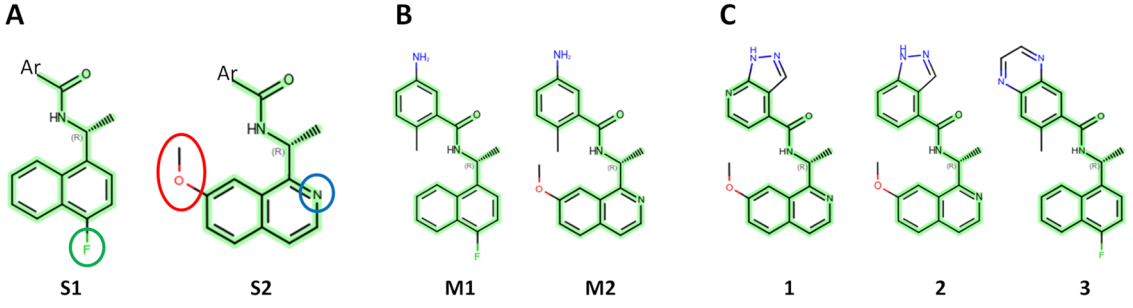


Figure 8. Ligands with improved binding affinity. (A) Common scaffolds of ligands with negative ΔΔ*G*. "Ar" corresponds to any substituted aromatic system containing a six-membered aromatic ring. Chemical modifications with respect to the scaffold of reference ligand are circled. (B) Reference ligand analogs corresponding to the common scaffolds shown in section A. (C) Ligands with the highest predicted binding affinity.

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

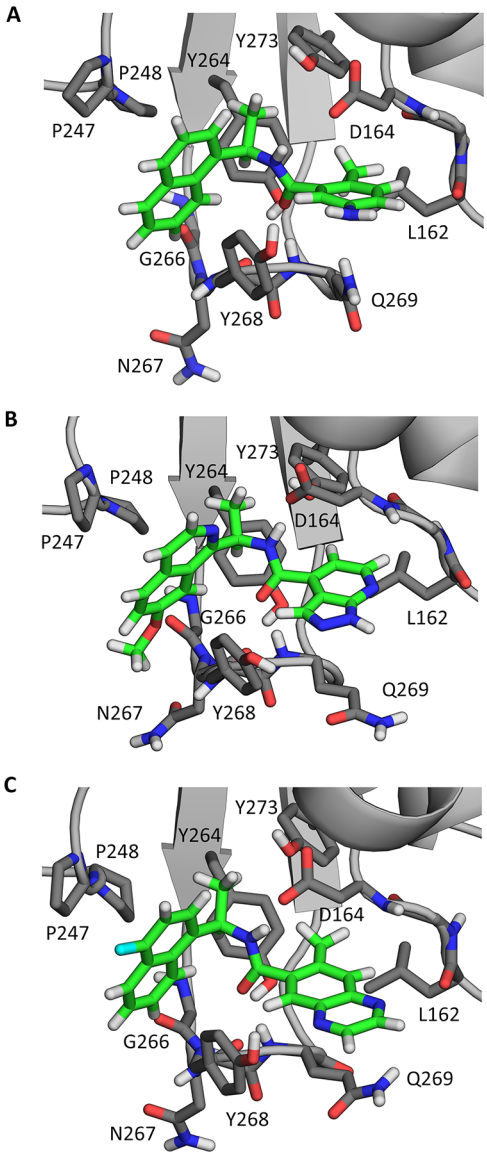


Figure 9. Representative binding poses of the reference ligand (A), ligand 1 (B), and ligand 3 (C). Carbon atoms of ligands and protein residues are shown in green and gray, respectively. Nitrogen, oxygen, and fluorine atoms are shown in blue, red, and cyan, respectively.

$\Delta pK_d = 2$ correspond to a 10-fold and 100-fold improvement in the binding affinity (K_d of 270 and 27 nM, respectively). Most ligands selected by the AL workflow (~ 70%) were found to have a stronger binding affinity ($\Delta pK_d > 0$). In particular, around 90% of the ligands selected by the AL at the exploitative phase have a stronger binding affinity than the reference ligand (Figure 6A). In contrast, most ligands selected randomly or by diversity (~89%) were found to have a weaker binding affinity than the reference ligand ($\Delta pK_d < 0$; Figure 6B). Remarkably, ~9% of ligands selected by the AL had $\Delta pK_d > 2$ while only one such ligand was found with random sampling. Overall, the distribution of predicted ΔpK_d for ligands selected by AL is substantially shifted toward higher binding affinity. Thus, these data show that employing AL results in a significant increase in the efficiency of the alchemical calculations for virtual screening.

In the exploitative cycle 7, 27 out of 30 ligands were found to have improved binding affinity with respect to the reference ligand. In contrast, in the random sample, the distribution was the inverse, with only 3 out of 30 ligands having an improved binding affinity.

TI RBFE calculations were performed for 253 ligands. Negative RBFEs were computed for 133 ligands, *i.e.*, approximately 53% of TI calculations. Thus, more than half of the ligands screened by the TI calculations were predicted to have higher binding affinity than the reference ligand. Among these, 62 ligands, or 24.5% of the ligands screened by the TI, were found to have more than a ten-fold improvement in predicted binding affinity. Given that the dissociation constant for the reference ligand was 2.7 μM,[17] dissociation constants for these ligands were predicted to be smaller than 270 nM. 16 ligands, or 6% of the ligands screened by TI, were found to have more than a hundred-fold improvement in predicted binding affinity to the target protein, which corresponds to dissociation constants below 27 nM.

Among ML-selected molecules (in the explorative and exploitative cycles), approximately 70% were estimated by TI to have higher binding affinity than the reference ligand. In contrast, the ratio of such ligands for the diverse and random samples was only about 10%. Notably, our results demonstrated a significant advantage of the Auto-ML guided sampling over the random and diverse sampling in identifying ligands with more than ten-fold improvement in predicted binding affinity. In the ML samples, approximately 25% and approximately 8% of ligands had 10−100 and greater than 100-fold improvement in binding affinity, respectively, while in random and diversity samples, both ratios were approximately 1%.

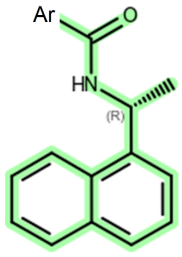
Figure 7 shows the evolution of the ML model's perception of the focused library chemical space, as well as the distribution of the molecules chosen for the TI RBFE calculations. In Figure 7, the focused library chemical space is depicted as a two-dimensional t-SNE projection, which estimates an organization of the high-dimensional representation of the molecular chemical space and constructs a low-dimensional representation that preserves relationships present in the high-dimensional representation.[52]

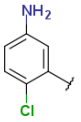
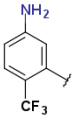
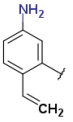
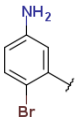
Notably, at the beginning of the active learning workflow (AL cycle 0), the ML model does not distinguish (Figure 7) specific regions of the chemical space enriched with favorable binders characterized by low ΔΔG. In the following AL cycles 1−5, with the balanced selection employed, the model was exploring multiple regions and found the perspective chemical space (Figure 7, AL 1−5, TI row). As a result of information gain, the ML model's perception was changing significantly (Figure 7, AL 0, ML row). Regions of chemical space that are densely populated with low $\Delta\Delta G$ molecules started to be identified (see Figure7, AL 1, ML row). By AL cycle 5, the ML model is converged (Figure 5), which is indicated by the stabilized coloring of various regions of the chemical space. During AL cycle 6 (Figure 7, AL 6, TI row), which employs a random selection of the molecules, sampled molecules are spread across the chemical space and, as expected, the majority of molecules have a positive $\Delta\Delta G$. Notably, the model's errors (Figure 5) did not increase, which supports the observation of model convergence. Thus, we conclude our study with the exploitative AL cycle 7, as discussed above.

Analysis of the Ligands with Improved Predicted Binding Affinity. Two common modifications in the naphthalene ring of *N*-[(1*R*)-1-arylethyl]arenecarboxamide were present in the molecules with improved predicted binding affinity (scaffolds

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

Table 1. Experimental and Computed Binding Free Energies for the Benchmark Set of PLpro Inhibitors[a]



No.	Ar	$\Delta\Delta G_{exp}$ (kcal/mol)	$\Delta\Delta G_{TI}$ (kcal/mol)	$\Delta\Delta G_{ML}$ (kcal/mol)	ΔG_{exp} (kcal/mol)	ΔG_{TI} (kcal/mol)
20		-0.18	0.14±0.06	1.25	-7.78	-7.46±0.06
21		3.03	1.01±0.09	1.72	-4.56	-6.59±0.09
22		1.78	1.50±0.09	1.63	-5.82	-6.10±0.09
23		2.02	0.14±0.06	3.20	-5.57	-7.45±0.06
27		2.98	2.25±0.13	2.88	-4.61	-5.35±0.13

[a]Ligand numbers (No.) are given in accordance with Shen et al.[17] ΔG_{exp} is the experimental absolute binding free energy converted from dissociation constant by eq 6 (see Methods). $\Delta\Delta G_{exp}$ is the experimental RBFE with respect to the reference ligand obtained from ΔG_{exp} by eq 7. $\Delta\Delta G_{TI}$ is the RBFE computed by TI. ΔG_{TI} is the absolute binding free energy obtained from $\Delta\Delta G_{TI}$ by eq 8. $\Delta\Delta G_{ML}$ is the RBFE predicted by the final ML model.

S1 and **S2** in Figure 8A). The first modification (**S1**) is a substitution of hydrogen by fluorine in position 4 of the naphthalene ring. The second modification (**S2**) includes substitution of the β-naphthalene carbon to nitrogen and an addition of a methoxy group in position 7 of the aromatic ring, which makes it a 7-methoxyisoquinoline moiety. To assess the relative importance of these modifications, we computed the RBFEs GRL0617 → **M1** and GRL0617 → **M2** (Figure 8B), which resulted in improved binding affinity by −0.84 and −0.99 kcal/mol, respectively. The third common structural feature of ligands with improved predicted binding affinity was the presence of fused 5,6- and 6,6-bicyclic aromatic systems in place of the benzene ring of the reference ligand (Figure 8C). Among ligands with negative TI $\Delta\Delta G$, there were 35 (∼26%) molecules with similar aromatic systems. Nine of these molecules showed more than a 100-fold improvement in predicted binding affinity (TI $\Delta\Delta G$ < 2.73 kcal/mol; see Figure S2). For ligands with the highest predicted $\Delta\Delta G$ (ligands **1**−**3**, see Figure 8C), computed TI $\Delta\Delta G$ were −4.06, −4.05, and −3.72 kcal/mol, respectively, which corresponds to dissociation constants of 2.85, 2.90, and 5.07 nM (these values were obtained by converting TI $\Delta\Delta G$ to ΔG by eq 8 and then converting these ΔG to K_d according to eq 6).

The reference ligand benzene ring substituents displayed specific interactions with the protein: the amino group formed hydrogen bonds with both the amide group of Gln269 and the hydroxyl group of Tyr268, and the methyl group formed hydrophobic interactions with the side chains of Tyr264, Tyr273, and Leu162 (Figure 9A). The representative binding poses of the ligands **1** and **3** and the reference ligand are shown in the Figure 9B,C. The amide group of the linker formed hydrogen bonds with the main-chain amino group of Gln269, the hydroxyl group of Tyr264, and the carboxylic group of Asp164. The naphthalene ring of the reference ligand and ligand **3** and isoquinoline ring of ligands **1** and **2** form hydrophobic interactions with the side chains of Pro248 and Tyr268. The benzene ring of the reference ligand and ligands **2**, **3**, and the pyridine ring of ligand **1** form hydrophobic interactions with aliphatic regions of the side chains of Gln269 and Asp164.

Modifications present in ligands **1**−**3** allow for several protein-ligand interactions absent in the reference ligand. The methoxy group of ligands **1** and **2** forms polar interactions with the main chains of Gly266 and Asn267. The pyrazole ring of ligands **1** and **2** and also the pyrazine ring of ligand **3** form polar interactions with the side chains of Tyr268 and Gln269. Notably, the analog of ligand **3**, in which a benzene ring methyl substituent is absent, has a TI $\Delta\Delta G$ of −1.56 kcal/mol, which suggests that the presence of this methyl group is important for binding affinity improvement.

Benchmarking the Free Energy Calculation Protocol. To validate the performance of the TI RBFE calculation protocol used in this work, we performed RBFE calculations for five PLpro inhibitors for which experimental binding affinities were reported by Shen et al.[17] (Table 1). Ligands for benchmarking were chosen to be non-protonatable at physiological pH. The same ligands' RBFEs were also predicted using the final ML model (Table 1).

For three ligands (**20**, **22**, and **27**), the absolute error between the experimental and the computed RBFEs is below 1

https://doi.org/10.1021/acs.jcim.2c01052
J. Chem. Inf. Model. 2023, 63, 583−594

kcal/mol. For ligands **21** and **23**, the absolute error is approximately 2 kcal/mol. In this case, the RBFE calculations underestimated a decrease in the binding affinity caused by the substitution of the benzene ring methyl substituent with a trifluoromethyl group (ligand **21**) or a bromine atom (ligand **23**). This discrepancy may be due to poorer parametrization of halogens compared to CHON atoms in GAFF.

The MAE of computed RBFEs with respect to the experimental values was 1.07 kcal/mol for TI and 0.83 kcal/mol for ML. The TI RBFE MAE was comparable with benchmarks reported recently.[8−11] In particular, MAEs for the RBFE calculations performed using the same lambda window schedule as used in this work were in the range of 0.74−0.91 kcal/mol for four protein−ligand systems;[11] however, absolute errors of 2 kcal/mol and above were reported for a small number of ligands in the same studies. Therefore, it is still challenging to achieve accurate RBFE predictions for all ligands included in a chemical space of interest. Further advances in force fields will allow for improved accuracy of the RBFE calculations.

CONCLUSIONS

Lead optimization remains a substantial computational challenge for modern computational chemistry. Computationally intensive campaigns, such as molecular dynamics for relative binding free energy simulations, are typically severely limited by the availability of computational resources as well as the difficulty of performing computations in a high-throughput manner. For example, the COVID-19 Moonshot initiative ran over 5000 free energy simulations exploiting the global Folding@home computational initiative.[53] This massive undertaking used hundreds of millions of computer hours to achieve a 100-fold improvement in potency against the SARS-CoV-2 main protease. Such resources are rarely available. Here, we were able to perform RBFE calculations only for a subset of ligands, rather than for all available analogs of a lead compound by coupling such calculations with an active learning approach, which included an automatic machine learning model selection.

Using a selection of molecules enriched by the Auto-ML procedure, we identified 133 potential SARS-CoV-2 PLpro inhibitors predicted to have improved binding affinity by performing the TI RBFE calculations for only 253 ligands. Remarkably, the alchemical RBFE calculations predicted improved binding affinity for 70% of ligands selected by ML in contrast to only 11% of ligands selected randomly. We believe that the approach developed here is an important step toward accelerating the lead optimization stage of drug design projects by leveraging modern computational approaches.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.jcim.2c01052.

Methods, additional details on alchemical free energy calculations and thermodynamic integration; Figure S1, distribution of the docked poses before and after filtering; Figure S2, chemical structures of molecules with $\Delta\Delta G$ values < −2.50 kcal/mol and the reference ligand; Table S1, list of molecular representations and featurization used for ML modeling; and Table S2, training data parameters and optimized AutoML parameters for each active learning cycle (PDF)

ZIP archive with docked poses for molecules used in TI calculations (ZIP)

AUTHOR INFORMATION

Corresponding Authors

Maria G. Kurnikova − *Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0002-8010-8374; Email: kurnikova@cmu.edu

Olexandr Isayev − *Department of Chemistry and Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0001-7581-8497; Email: olexandr@olexandrisayev.com

Authors

Filipp Gusev − *Department of Chemistry and Computational Biology Department, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0002-1167-345X

Evgeny Gutkin − *Department of Chemistry, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, United States;* orcid.org/0000-0003-4522-6049

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jcim.2c01052

Author Contributions

[§]These authors contributed equally and share the first authorship.

Notes

The authors declare no competing financial interest.

Molecular docking was performed using the OpenEye HYBRID (version 4.0.0.0) program.[22] Atom mappings were generated by the LS-align tool.[23] Ligand alignment and system setup for TI RBFE simulations were performed using the FESetup tool (version 1.2.1).[32] Conventional and TI RBFE simulations were performed using the pmemd.cuda module of the AMBER 18 package.[25] Initial poses for all ligands for which TI RBFE simulations were performed are available in the Supporting Information. Molecular representations were constructed using RDKit software (https://www.rdkit.org/, version 2020.09.5). AutoML was implemented using the scikit-learn package[38,50] (version 0.23.2). All parameters of docking, MD simulations, and ML are described in the Methods section.

ACKNOWLEDGMENTS

We acknowledge support from DSF Charitable Foundation and the COVID-19 HPC Consortium. O.I. research is supported by grants NSF CHE-2154447 and CHE-2041108, and M.K. is supported by grants NSF DMS-1563291, MCB-1818213, and NIH R01NS083660. The authors acknowledge Extreme Science and Engineering Discovery Environment (XSEDE) supported by NSF ACI-1053575 and Frontera computing project at the Texas Advanced Computing Center (NSF OAC-1818253) award.

REFERENCES

(1) Jorgensen, W. L., Progress and Issues for Computationally Guided Lead Discovery and Optimization. In *Drug Design: Structure-*

and Ligand-Based Approaches; Reynolds, C. H.; Ringe, D.; Merz, J. K. M., Eds. Cambridge University Press: Cambridge, 2010; pp. 1−14, DOI: 10.1017/CBO9780511730412.003.
(2) Steinbrecher, T. Free Energy Calculations in Drug Lead Optimization. *Protein-Ligand Interact.* **2012**, 207−236.
(3) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23*, 101681.
(4) Sadybekov, A. A.; Sadybekov, A. V.; Liu, Y.; Iliopoulos-Tsoutsouvas, C.; Huang, X.-P.; Pickett, J.; Houser, B.; Patel, N.; Tran, N. K.; Tong, F.; Zvonok, N.; Jain, M. K.; Savych, O.; Radchenko, D. S.; Nikas, S. P.; Petasis, N. A.; Moroz, Y. S.; Roth, B. L.; Makriyannis, A.; Katritch, V. Synthon-Based Ligand Discovery in Virtual Libraries of over 11 Billion Compounds. *Nature* **2022**, *601*, 452−459.
(5) Lyu, J.; Wang, S.; Balius, T. E.; Singh, I.; Levit, A.; Moroz, Y. S.; O'Meara, M. J.; Che, T.; Algaa, E.; Tolmachova, K.; Tolmachev, A. A.; Shoichet, B. K.; Roth, B. L.; Irwin, J. J. Ultra-Large Library Docking for Discovering New Chemotypes. *Nature* **2019**, *566*, 224−229.
(6) Kirkwood, J. G. Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **1935**, *3*, 300−313.
(7) Mey, A. S. J. S.; Allen, B. K.; Bruce McDonald, H. E.; Chodera, J. D.; Hahn, D. F.; Kuhn, M.; Michel, J.; Mobley, D. L.; Naden, L. N.; Prasad, S.; Rizzi, A.; Scheen, J.; Shirts, M. R.; Tresadern, G.; Xu, H. Best Practices for Alchemical Free Energy Calculations [Article V1.0]. *Living J. Comp. Mol. Sci* **2020**, *2*, 18378.
(8) Wang, L.; Wu, Y.; Deng, Y.; Kim, B.; Pierce, L.; Krilov, G.; Lupyan, D.; Robinson, S.; Dahlgren, M. K.; Greenwood, J.; Romero, D. L.; Masse, C.; Knight, J. L.; Steinbrecher, T.; Beuming, T.; Damm, W.; Harder, E.; Sherman, W.; Brewer, M.; Wester, R.; Murcko, M.; Frye, L.; Farid, R.; Lin, T.; Mobley, D. L.; Jorgensen, W. L.; Berne, B. J.; Friesner, R. A.; Abel, R. Accurate and Reliable Prediction of Relative Ligand Binding Potency in Prospective Drug Discovery by Way of a Modern Free-Energy Calculation Protocol and Force Field. *J. Am. Chem. Soc.* **2015**, *137*, 2695−2703.
(9) Song, L. F.; Lee, T.-S.; Zhu, C.; York, D. M.; Merz, K. M. Using AMBER18 for Relative Free Energy Calculations. *J. Chem. Inf. Model.* **2019**, *59*, 3128−3135.
(10) Kuhn, M.; Firth-Clark, S.; Tosco, P.; Mey, A. S. J. S.; Mackey, M.; Michel, J. Assessment of Binding Affinity Via Alchemical Free-Energy Calculations. *J. Chem. Inf. Model.* **2020**, *60*, 3120−3130.
(11) He, X.; Liu, S.; Lee, T. S.; Ji, B.; Man, V. H.; York, D. M.; Wang, J. Fast, Accurate, and Reliable Protocols for Routine Calculations of Protein-Ligand Binding Affinities in Drug Design Projects Using AMBER GPU-TI with ff14SB/GAFF. *ACS Omega* **2020**, *5*, 4611−4619.
(12) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations to Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *J. Chem. Inf. Model.* **2019**, *59*, 3782−3793.
(13) Lee, T.-S.; Allen, B. K.; Giese, T. J.; Guo, Z.; Li, P.; Lin, C.; McGee, T. D.; Pearlman, D. A.; Radak, B. K.; Tao, Y.; Tsai, H.-C.; Xu, H.; Sherman, W.; York, D. M. Alchemical Binding Free Energy Calculations in AMBER20: Advances and Best Practices for Drug Discovery. *J. Chem. Inf. Model.* **2020**, 5595.
(14) Abel, R.; Wang, L.; Mobley, L. D.; Friesner, A. R. A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Curr. Top. Med. Chem.* **2017**, *17*, 2577−2585.
(15) Shin, D.; Mukherjee, R.; Grewe, D.; Bojkova, D.; Baek, K.; Bhattacharya, A.; Schulz, L.; Widera, M.; Mehdipour, A. R.; Tascher, G.; Geurink, P. P.; Wilhelm, A.; van der Heden van Noort, G. J.; Ovaa, H.; Müller, S.; Knobeloch, K.-P.; Rajalingam, K.; Schulman, B. A.; Cinatl, J.; Hummer, G.; Ciesek, S.; Dikic, I. Papain-Like Protease Regulates SARS-Cov-2 Viral Spread and Innate Immunity. *Nature* **2020**, *587*, 657−662.
(16) Osipiuk, J.; Azizi, S.-A.; Dvorkin, S.; Endres, M.; Jedrzejczak, R.; Jones, K. A.; Kang, S.; Kathayat, R. S.; Kim, Y.; Lisnyak, V. G.; Maki, S. L.; Nicolaescu, V.; Taylor, C. A.; Tesar, C.; Zhang, Y.-A.; Zhou, Z.; Randall, G.; Michalska, K.; Snyder, S. A.; Dickinson, B. C.; Joachimiak, A. Structure of Papain-Like Protease from SARS-Cov-2 and Its Complexes with Non-Covalent Inhibitors. *Nat. Commun.* **2021**, *12*, 743.
(17) Shen, Z.; Ratia, K.; Cooper, L.; Kong, D.; Lee, H.; Kwon, Y.; Li, Y.; Alqarni, S.; Huang, F.; Dubrovskyi, O.; Rong, L.; Thatcher, G. R. J.; Xiong, R. Design of SARS-Cov-2 PLpro Inhibitors for Covid-19 Antiviral Therapy Leveraging Binding Cooperativity. *J. Med. Chem.* **2022**, *65*, 2940−2955.
(18) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887−2893.
(19) *OpenEye Toolkits 2021.2.0 OpenEye Scientific Software*, Santa Fe, NM. http://www.eyesopen.com
(20) *OMEGA 4.1.0.0: OpenEye Scientific Software*, Santa Fe, NM. http://www.eyesopen.com
(21) *Make Receptor 4.0.0.0: OpenEye Scientific Software, Inc.*, Santa Fe, NM. http://www.eyesopen.com
(22) McGann, M. FRED and HYBRID Docking Performance on Standardized Datasets. *J. Comput.-Aided Mol. Des.* **2012**, *26*, 897−906.
(23) Hu, J.; Liu, Z.; Yu, D.-J.; Zhang, Y. LS-align: An Atom-Level, Flexible Ligand Structural Alignment Algorithm for High-Throughput Virtual Screening. *Bioinformatics* **2018**, *34*, 2209−2218.
(24) Osipiuk, J.; Tesar, C.; Endres, M.; Lisnyak, V.; Maki, S.; Taylor, C.; Zhang, Y.; Zhou, Z.; Azizi, S.A.; Jones, K.; Kathayat, R.; Snyder, S.A.; Dickinson, B.C.; Joachimiak, A., *Center for Structural Genomics of Infectious Diseases (CSGID) The Crystal Structure of Papain-Like Protease of SARS Cov-2 , C111s Mutant, in Complex with PLP_Snyder457 Inhibitor*. 2020, DOI: 10.2210/pdb7jir/pdb.
(25) Case, D. A.; Ben-Shalom, I. Y.; Brozell, S. R.; Cerutti, D. S.; Cheatham, T. E.; Iii, V. W. D.; Cruzeiro, T. A.; Darden, R. E.; Duke, D.; Ghoreishi, M. K.; Gilson, H.; Gohlke, A. W.; Goetz, D. G.; Harris, R.; Homeyer, N.; Huang, Y.; Izadi, S.; Kovalenko, A.; Kurtzman, T.; Lee, T. S.; Legrand, S.; Li, P.; Lin, C.; Liu, J.; Luchko, T.; Luo, R.; Mermelstein, D. J.; Merz, K. M.; Miao, Y.; Monard, G.; Nguyen, C.; Nguyen, H.; Omelyan, I.; Onufriev, A.; Pan, F.; Qi, R.; Roe, D. R.; Roitberg, A.; Sagui, C.; Schott-Verdugo, S.; Shen, J.; Simmerling, C. L.; Smith, J.; Salomonferrer, R.; Swails, J.; Walker, R. C.; Wang, J.; Wei, H.; Wolf, R. M.; Wu, X.; Xiao, L.; York, D. M.; Kollman, P. A.*AMBER 2018*; University of California, San Francisco2018.
(26) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from Ff99sb. *J. Chem. Theory Comput.* **2015**, *11*, 3696−3713.
(27) Peters, M. B.; Yang, Y.; Wang, B.; Füsti-Molnár, L.; Weaver, M. N.; Merz, K. M. Structural Survey of Zinc-Containing Proteins and Development of the Zinc AMBER Force Field (ZAFF). *J. Chem. Theory Comput.* **2010**, *6*, 2935−2947.
(28) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. Comparison of Simple Potential Functions for Simulating Liquid Water. *J. Chem. Phys.* **1983**, *79*, 926−935.
(29) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and Testing of a General Amber Force Field. *J. Comput. Chem.* **2004**, *25*, 1157−1174.
(30) Jakalian, A.; Jack, D. B.; Bayly, C. I. Fast, Efficient Generation of High-Quality Atomic Charges. AM1-BCC Model: II. Parameterization and Validation. *J. Comput. Chem.* **2002**, *23*, 1623−1641.
(31) Shirts, M. R.; Mobley, D. L., An Introduction to Best Practices in Free Energy Calculations. In *Biomolecular Simulations: Methods and Protocols*, Monticelli, L.; Salonen, E., Eds. Humana Press: Totowa, NJ, 2013; pp. 271−311.
(32) Loeffler, H. H.; Michel, J.; Woods, C. Fesetup: Automating Setup for Alchemical Free Energy Simulations. *J. Chem. Inf. Model.* **2015**, *55*, 2485−2490.
(33) *The PyMOL Molecular Graphics System, Version 1.8.4.0*, Schrödinger, LLC.

(34) Boresch, S.; Tettinger, F.; Leitgeb, M.; Karplus, M. Absolute Binding Free Energies: A Quantitative Approach for Their Calculation. *J. Phys. Chem. B* **2003**, *107*, 9535−9551.
(35) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. *J. Chem. Theory Comput.* **2016**, *12*, 1799−1805.
(36) Axen, S. D.; Huang, X.-P.; Cáceres, E. L.; Gendelev, L.; Roth, B. L.; Keiser, M. J. A Simple Representation of Three-Dimensional Molecular Structure. *J. Med. Chem.* **2017**, *60*, 7393−7409.
(37) Wójcikowski, M.; Kukiełka, M.; Stepniewska-Dziubinska, M. M.; Siedlecki, P. Development of a Protein−Ligand Extended Connectivity (PLEC) Fingerprint and Its Application for Binding Affinity Predictions. *Bioinformatics* **2019**, *35*, 1334−1341.
(38) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825−2830.
(39) Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way to Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 747−750.
(40) Settles, B.*Active Learning Literature Survey*. 2009.
(41) Ambroise, C.; McLachlan, G. J. Selection Bias in Gene Extraction on the Basis of Microarray Gene-Expression Data. *Proc. Natl. Acad. Sci. U. S. A.* **2002**, *99*, 6562.
(42) Ahneman Derek, T.; Estrada Jesús, G.; Lin, S.; Dreher Spencer, D.; Doyle Abigail, G. Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning. *Science* **2018**, *360*, 186−190.
(43) Chuang Kangway, V.; Keiser Michael, J. Comment on "Predicting Reaction Performance in C−N Cross-Coupling Using Machine Learning.". *Science* **2018**, *362*, eaat8603.
(44) DeVries, P. M. R.; Viégas, F.; Wattenberg, M.; Meade, B. J. Deep Learning of Aftershock Patterns Following large Earthquakes. *Nature* **2018**, *560*, 632−634.
(45) Mignan, A.; Broccardo, M. One Neuron Versus Deep Learning in Aftershock Prediction. *Nature* **2019**, *574*, E1−E3.
(46) He, X.; Zhao, K.; Chu, X. Automl: A Survey of the State-of-the-Art. *Knowledge-Based Syst.* **2021**, *212*, 106622.
(47) Tuggener, L.; Amirian, M.; Rombach, K.; Lörwald, S.; Varlet, A.; Westermann, C.; Stadelmann, T. In *Automated Machine Learning in Practice: State of the Art and Recent Results*, 2019 6th Swiss Conference on Data Science (SDS), 14−14 June 2019; 2019; pp. 31−36.
(48) Chauhan, K.; Jani, S.; Thakkar, D.; Dave, R.; Bhatia, J.; Tanwar, S.; Obaidat, M. S. In *Automated Machine Learning: The New Wave of Machine Learning*, 2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 5−7 March 2020; 2020; pp. 205−212.
(49) Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J.; Blum, M.; Hutter, F.Efficient and Robust Automated Machine Learning. *Adv. Neural Inf.. Proc. Syst.***2015**, *28*.
(50) Feurer, M.; Klein, A.; Eggensperger, K.; Springenberg, J. T.; Blum, M.; Hutter, F., Auto-Sklearn: Efficient and Robust Automated Machine Learning. In *Automated Machine Learning: Methods, Systems, Challenges*, Hutter, F.; Kotthoff, L.; Vanschoren, J., Eds. Springer International Publishing: Cham, 2019; pp. 113−134, DOI: 10.1007/978-3-030-05318-5_6.
(51) Mann, H. B.; Whitney, D. R. On a Test of Whether One of Two Random Variables Is Stochastically Larger Than the Other. *Annals Math Stat.* **1947**, *18*, 50−60 11.
(52) Van der Maaten, L.; Hinton, G.Visualizing Data Using t-SNE. *J. Mach. Learn. Res.***2008**, *9*.
(53) von Delft, F.; Calmiano, M.; Chodera, J.; Griffen, E.; Lee, A.; London, N.; Matviuk, T.; Perry, B.; Robinson, M.; von Delft, A. A White-Knuckle Ride of Open Covid Drug Discovery. *Nature* **2021**, *594*, 330−332.