Active learning guided drug design lead optimization based on relative binding free energy modeling

Filipp Gusev[1,2,†], Evgeny Gutkin[1,†], Maria G. Kurnikova[1*], Olexandr Isayev[1,2*]

[1]*Chemistry Department, Carnegie Mellon University, Pittsburgh, PA 15213*

[2]*Computational Biology Department, Carnegie Mellon University, Pittsburgh, PA, 15213*

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

* Corresponding authors:

Maria G. Kurnikova
Email: kurnikova@cmu.edu

Olexandr Isayev
Email: olexandr@olexandrisayev.com

Author Contributions: F.G. performed dataset preparation and machine learning. E.G. performed molecular docking and molecular dynamics simulations. All authors designed research, analyzed results and wrote the manuscript.

Competing Interest Statement: The authors declare no competing interests.

Classification: Physical Sciences/Biophysics and Computational Biology, Physical Sciences/Computer Sciences

Keywords: Molecular dynamics, free energy calculations, machine learning, active learning, COVID-19, SARS-CoV-2 PLpro

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 predicting accurate BFE, but it is computationally expensive and time-consuming. In this work, we 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 AL-AutoML workflow allows unbiased and efficient search for a small set of best performing molecules. We applied this workflow to select inhibitors of the SARS-CoV-2 papain-like protease. Our work resulted in predicting 133 compounds with improved binding affinity among which 16 compounds with better than 100-fold binding affinity improvement. The hit rate obtained here is better than that of traditional projects where molecule selection is guided by an expert medicinal chemist. We demonstrated that a combination of an AL protocol provides at least 20x the common brute force approaches.

Significance Statement

Hit-to-lead and lead optimization stages of drug discovery have low success rates. In silico screening of ultra-large molecular libraries is an attractive alternative to experimental approaches and expert-driven prediction, yet current success of virtual screening is limited by poor predictive power of ligand ranking methods. Incorporation of molecular dynamics (MD) for computing relative binding free energy (RBFE) provides a solution. However, MD methodologies are restricted by intensive computing resource requirements. To efficiently overcome such restraints we developed a workflow which relies on Automated Machine Learning guided by Active Learning approach to minimize a number of required MD-RBFE evaluations.

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 chemical structure of a hit molecule that has a demonstrated activity against the target. A typical process of lead optimization involves expensive and time-consuming process of first, chemical synthesis of multiple compounds and then, testing them for biological activity (1, 2). Structure-based virtual screening of ultra-large molecular libraries aimed at minimizing the number of compounds chosen for laboratory synthesis and testing has become a successful strategy in computational drug design (3). High hit rates were achieved with docking ligands to target protein (4). However, two main limitations of

such approaches remain: a limited ability of docking methodologies to predict ligand binding affinity, and technological difficulty in working with libraries comprised of multibillion 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 a hit-to-lead and lead optimization (8-12). However, despite recent advances in high performance computing and improvement of algorithms for graphical processing units (GPU)-accelerated MD simulations, computing multiple RBFEs for a large number of compounds remains prohibitively time consuming and technically intractable (13).

To overcome this problem, we developed an automated approach for a machine learning (ML)-active learning (AL) guided lead optimization based on TI computed RBFEs. 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, 2) to improve an ML model 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 – MD TI RBFE computational workflow allows for identification of tight-binding ligands with a minimal number of the TI RBFE calculations.

The pandemics of a coronavirus disease 2019 (COVID-19) caused by a severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) remains a serious threat for global public health. 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 attractive drug targets for designing COVID-19 antivirals is SARS-CoV-2 papain-like protease (PLpro), an enzyme responsible for processing the viral polyprotein and suppressing the host immune function (14). 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 Fig.1). A fingers subdomain includes a zinc binding site formed by four cysteine residues. Protein active site is formed by a canonical cysteine protease catalytic triad which includes Cys111, His271 and Asp286 residues located at an interface between a thumb and a palm subdomains.

Several N-[(1R)-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 (15). In particular,

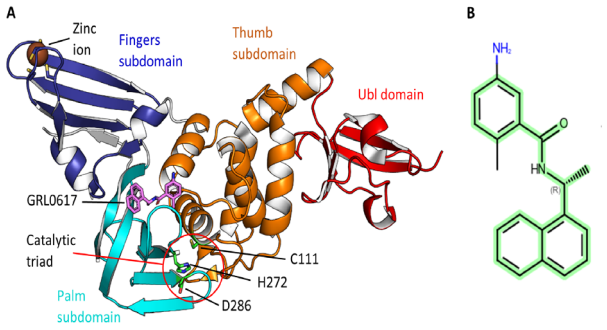

Figure. 1. Structure of SARS-CoV-2 PLpro and its inhibitor. A) Structure of SARS-CoV-2 PLpro in complex with GRL0617 (PDB ID: 7JIR). The inhibitor and residues of the catalytic triad are shown as pink and green sticks correspondingly. B) GRL0617 The common scaffold (N-[(1R)-1-arylethyl]arenecarboxamide)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

the most potent inhibitor GRL0617 with known high resolution structure (Fig. 1) demonstrates half maximal inhibitory concentration (IC_{50}) of 2.3 μM (15). 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 (15). 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 similar scaffold were proposed with limited success using an expert driven lead optimization approaches (16). In this work we virtually screened a library of 1.3 billion commercially available compounds, selected a focused library of ten thousand derivatives of N-[(1R)-1-arylethyl]arenecarboxamide, and finally identified sixteen potent binders with more than hundred-fold improvement in predicted binding affinity.

Results and discussion

Approach

The workflow developed in this work uses 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 as a source of training data for the ML models. To develop a focused library of compounds we initially screened a 1.3 billion of commercially available molecules from three reputable compound vendors Enamine, WuXi and Mcule. The resulting library of an approximately ten thousand N-[(1R)-1-arylethyl]arenecarboxamide derivatives was further narrowed down to 8175 compounds, which passed structural molecular docking quality control. These compounds were used to prepare bound poses by docking each molecule to the target binding site (see SI Methods, molecular docking subsection).

To find best PLpro binders we utilized an active learning (AL) approach. The AL was organized as an iterative cycle: i) starting with a seed set of molecules we perform TI RBFE calculations to obtain training data and train initial ML model, ii) we then select molecules for the next round of the TI RBFE calculations using current ML model, iii) we compute additional TI RBFEs for the molecules selected in ii) and re-train the ML model with an updated TI RBFE dataset. The cycle is 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. (17). Here AL workflow is organized as a black box optimization of $\Delta\Delta G$ obtained by the TI MD calculations for a subset of molecules from a focused library of molecules. The AL cycles are performed in two regimes: explorative and exploitative. These are distinguished by the data selection style: the explorative regime uses a balanced selection, while the exploitative regime uses a greedy selection (see SI Methods). The explorative regime is used until the ML model reaches an apparent convergence, followed by the exploitative regime used to select molecules with the lowest $\Delta\Delta G$ of binding. Both explorative and exploitative regimes are organized in an AL cycle of four steps (Fig. 2): (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 optimal set of candidate molecules for the TI MD calculations of RBFE, (4) perform the TI MD 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 (Fig. 2): 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 with the PLpro protein. The AL cycle shown in Fig 2 is initialized with a small but diverse set of molecules. Their $\Delta\Delta G$ values are computed with 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 MD calculation (see SI Methods for details). An ML model trained in Step 2 is used in the Step 3 to virtually screen chemical space (grey dots represents chemical space) to obtain ML-model

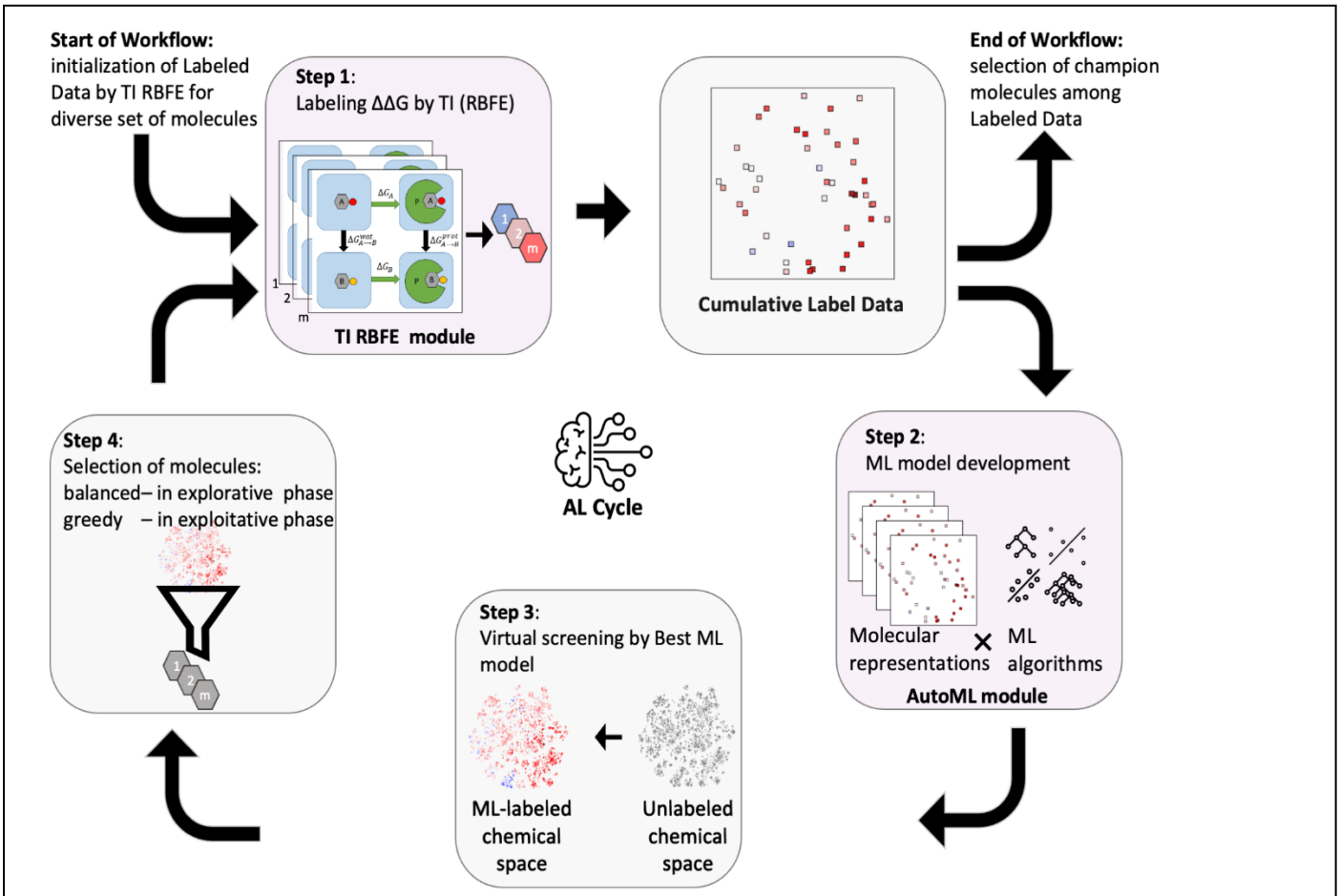

Figure 2. A general scheme of the automated computational workflow organized in an Active Learning Cycle. The workflow includes two main modules: AutoML and TI RBFE and four principal steps. Molecules with computed $\Delta\Delta G$ are depicted as colored hexagons in Step 1. Labeled chemical space is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

predicted $\Delta\Delta G$ values (color coded here as in Fig. 4). A new set of molecules is selected in Step 4 by the AL criteria to be submitted for the TI RBFE calculations thus completing the cycle.

Automated machine learning (AutoML) module (Fig. 2). Building an ML model with 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 substantial model bias, and a sample selection bias. Multiple studies showed that this bias may result in substantial modeling artifacts (18-22). In contrast, an AutoML aims to make decisions for an ML model selection, data representation and hyper-parameters in a data-driven, objective, and automated way (23-26). The combination of AutoML and AL approaches (AutoML-AL) allows for a fast systematic unbiased exploration of the chemical space in the first regime, and a selection of champion candidate molecules in the second, exploitative, regime. We implemented AutoML as a set of well performing ML algorithms available in scikit-learn package (27, 28), multiple molecular representation and on-the-fly feature engineering (see SI Methods for details).

Thermodynamic Integration MD for RBFE (TI RBFE). An automated protocol for the multiple RBFE calculations implemented in this work requires minimal user interaction. Our protocol accepts a set of docked ligands (see SI Methods section for details) as an input and provides calculated RBFEs for all ligands as an output. Compound GRL0617 (Fig. 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), and 2) set up and submission of the parallelized GPU-accelerated MD simulations using TI implementation of AMBER 18 package (29), 3) collection and processing of the output data. The details of the protocol are described in the SI Methods.

Results

An AutoML-AL approach was applied to perform eight AL cycles. Figure 3 shows all TI obtained ΔΔGs over all AL cycles. AL Cycle 0 was initialized with a diverse selection of molecules to sample chemical space of the focused library as wide as possible (see SI Methods for details). TI MD ΔΔG 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 with the low ML predicted RBFEs from clusters on a full focused library (see SI Methods 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 progression of AL cycles the performance of the ML model improved. The cross-validated mean absolute error (MAE; Fig. 3, 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 the molecules (see SI Methods). The random selection of molecules also serves to overcome a possible problem if AL being trapped in a local minimum of the chemical space.

We monitored two criteria between each balanced cycle (AL cycles 1-5) and a random cycle (AL cycle 6). First, the difference between the mean $\Delta\Delta Gs$ and second is retrospective MAE. The former is staying different up to ca. 2 kcal/mol. The latter remains nearly constant (Fig. 3, bottom) between the last balaced cycle (AL cycle 5) and the random cycle (AL cycle 6). This suggests that 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 $\Delta\Delta Gs$. The resulting AL Cycle 7 had mean TI ΔΔG with -1.7 kcal/mol as opposed to 2 kcal/mol in the AL Cycle 6 (Fig. 3, top). This difference is statistically significant with the p-value = $1.3\text{x}10^{-8}$ according to the Mann–Whitney U test (30).

The efficiency of the AL can be further demonstrated by comparison of the binding affinity distribution in the ligand samples selected by the ML models (including both the explorative and the exploitative sets), and in the molecular samples selected by diversity or randomly (see Fig. S1 and SI text). 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 inverse, with only 3 out of 30 ligands have an improved binding affinity.

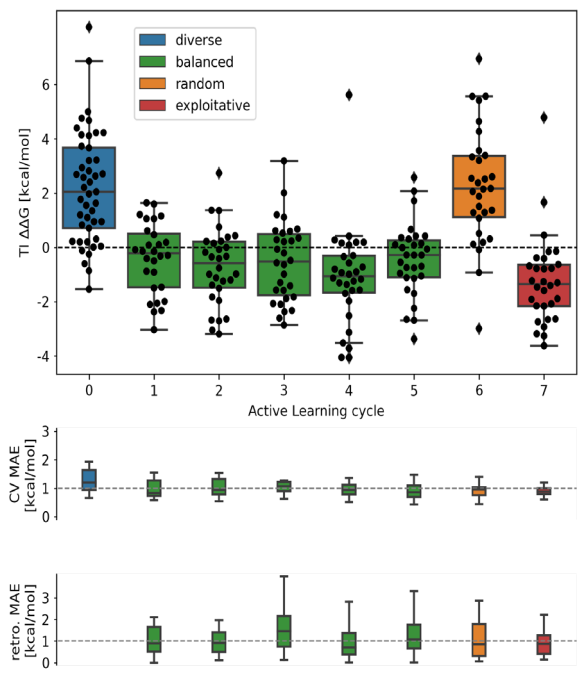

Figure 3. MD TI RBFE results in AL cycles (top); MAE for am ML model measured as 10-FOLD-CV (middle); retrospective MAE (bottom).

TI MD RBFE calculations were performed for 253 ligands. Negative RBFE were computed for 133 ligands, *i.e.* approximately 53% of TI calculations. Thus, more than half of the ligands screened by the TI MD 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 MD were found to have more than ten-fold improvement in predicted binding affinity. 16 ligands, or 6 % of the ligands screened by TI MD were found to have more than a hundred-fold improvement in predicted binding affinity to the target protein.

Among ML-selected molecules (in the explorative and exploitative cycles) approximately 70% were estimated by TI MD 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 ~25% and ~8% of ligands had 10-100 and > 100-fold improvement in binding affinity correspondingly while in random and diversity samples the both ratios are ~1%.

Figure 4 shows evolution of ML-model's perception of the chemical space of the focused library,

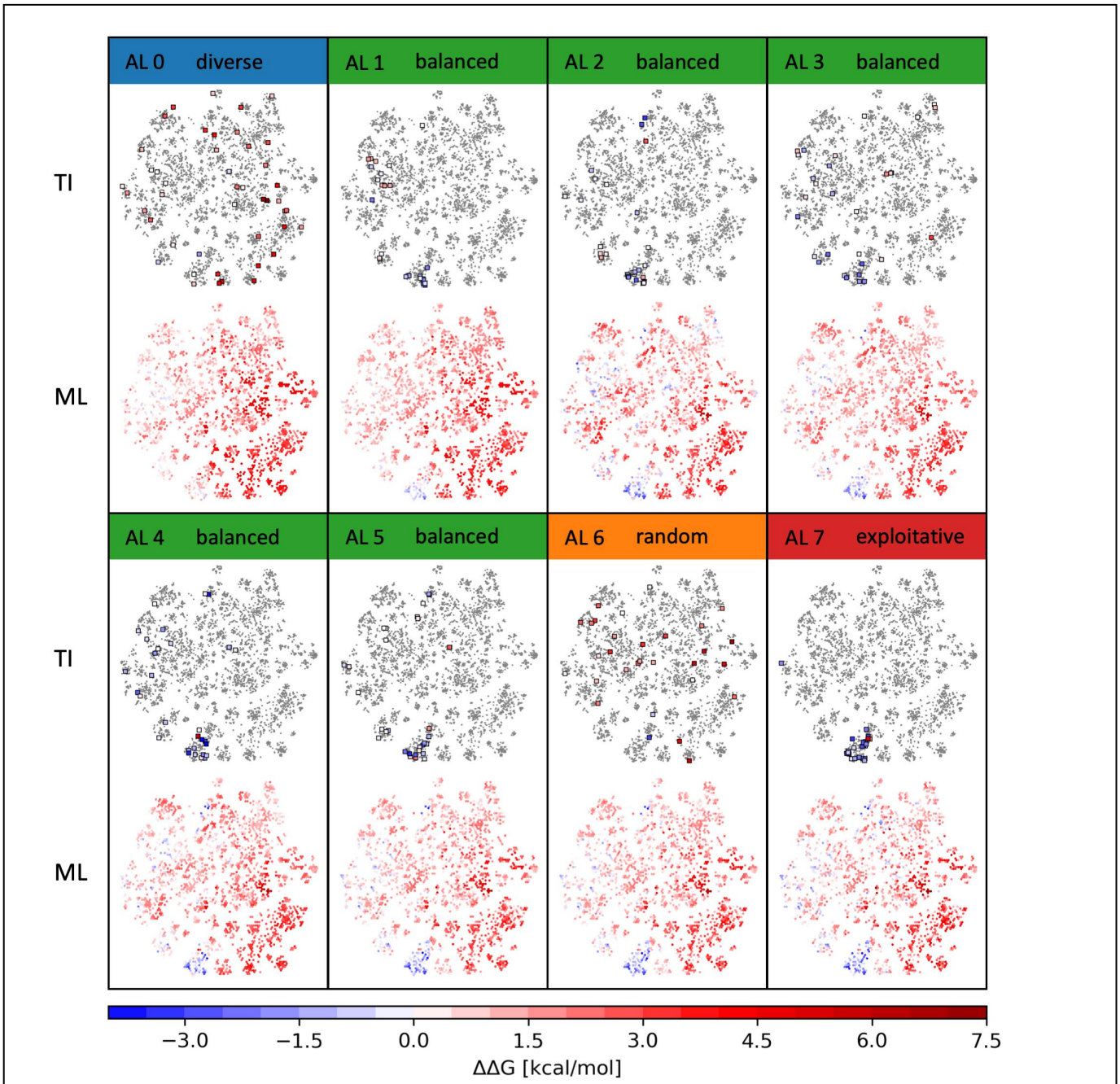

Figure 4. TI RBFE results and ML model evolution over the active learning cycles. Each panel (labeled by the AL cycle number and a corresponding selection style) shows two 2D labeled t-SNE representations of the focused library: (top) The molecules selected in arespective AL cycle for the TI RBFE calculation are colored by the TI computed $\Delta\Delta G$, the rest is shown in grey; (bottom) The focused library is colored by ML predicted $\Delta\Delta G$. Color bar for the $\Delta\Delta G$ values is shown at the bottom.

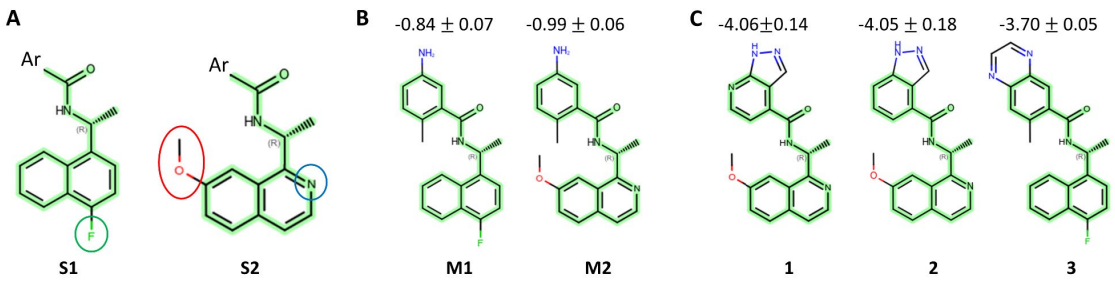

Figure 5. Ligands with improved binding affinity. (A) Common scaffolds of ligands with negative $\Delta\Delta G$. "Ar" corresponds to any substituted aromatic system containing a six-membered aromatic ring. Chemical modifications with respect to the scaffold of reference ligand are encircled. (B) Reference ligand analogs corresponding to the common scaffolds shown in section A. (C) Ligands with the highest predicted binding affinity. Calculated TI $\Delta\Delta G$ for ligands are shown above their structural formulas.

as well as a distribution of the molecules chosen for the TI MD RBFE calculations. In Fig. 4 the chemical space of the focused library 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 (31).

Notably, in the beginning of the active learning workflow, (the AL cycle 0) the ML-model does not distinguish (Fig. 4) 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, the model is exploring multiple regions and finding the perspective chemical space (Fig. 4, AL 1-5, TI row). As a result of information gain, the ML-model's perception is changing significantly (Fig. 4, AL 0, ML row). Regions of chemical space densely populated with low ΔΔG molecules started to be identified (See Fig. 4, AL 1, ML row). By the AL cycle 5 ML-model is converged (Fig. 3) which reflected in stabilized coloring of various regions of the chemical space. During the AL cycle 6 (Fig. 4, AL 6, TI row) with random selection the molecules, they are spread across the chemical space and majority of them have positive ΔΔG as expected. Notably, the model's errors (Fig. 3) did not increase, which supports an observation of model convergence. Thus, our study is concluded with an AL cycle 7 in an exploitative phase as discussed above.

Analysis of the ligands with improved predicted binding affinity. Two common modifications in the naphthalene ring of (N-[(1R)-1-arylethyl]arenecarboxamide) were present in the molecules with improved predicted binding affinity (scaffolds **S1** and **S2** in Fig. 5A). The first modification **(S1)** is a substitution in position 4 of the naphthalene ring. The second modification (**S2**) includes a substitution of the beta naphthalene carbon to nitrogen, and an addition of a methoxy group in position 7 of the aromatic ring, which make it a 7-methoxyisoquinoline moiety. To assess relative importance of these modifications we computed RBFEs. GRL0617 → **M1** and GRL0617 → **M2** (Fig. 5B) resulting in improved binding affinity by -0.84 kcal/mol 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. (Fig. 5C). Among ligands with negative TI $\Delta\Delta G$ there were 35 (~26%) molecules with similar aromatic systems. Nine of these molecules showed more than 100 fold improvement in predicted binding affinity (Fig. S2).

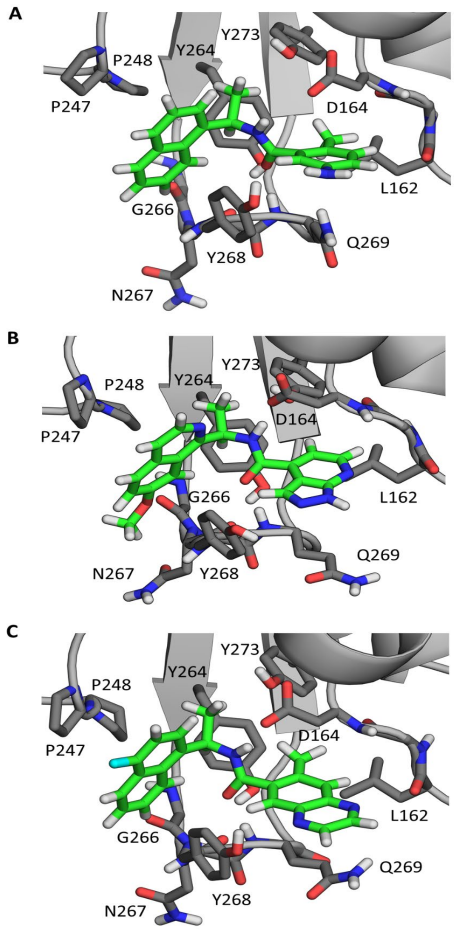

Figure 6. Representative binding poses of the reference ligand (A), ligand 1 (B) and ligand 3 (C).. Carbon atoms of ligands and protein residues are shown in green and grey correspondingly. Nitrogen, oxygen and fluorine atoms are shown in blue, red and cyan respectively.

The reference ligand has specific interactions with protein formed by benzene ring substituents: amino group makes hydrogen bond with amide group of Gln269 and hydroxyl group of Tyr268 and methyl group makes hydrophobic interactions with side chains of Tyr264, Tyr273 and Leu162 (Fig. 6A). The representative binding poses of the ligands **1** and **3** and the reference ligand are shown in the Fig. 6B and 6C. The amide group of the linker forms hydrogen bonds with the main-chain amino group of Gln269, hydroxyl group of Tyr264 and carboxylic group of Asp164. The naphthalene ring of the reference ligand and ligand **3** and isoquinoline ring of ligands **1**-**2** make hydrophobic interactions with side chains of Pro248 and Tyr268. Benzene ring of the reference ligand and ligands **2**-**3** and pyridine ring of ligand **1** make hydrophobic interactions with aliphatic regions of the side chains of Gln269 and Asp164.

Modifications present in ligands **1**-**3** allows for several protein-ligand interactions absent for the reference ligand. Methoxy group of ligands **1**-**2** makes polar interactions with main chains of Gly266 and Asn267. Pyrazole ring of ligands **1**-**2** and pyrazine ring of ligand **3** make polar interactions with side chains of Tyr268 and Gln269. Notably, the analog of ligand **3** in which methyl group of the benzene ring is absent has TI $\Delta\Delta G$ of -1.56 kcal/mol which suggests that the presence of this methyl group is important for binding affinity improvement.

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 availability of computational resources as well as difficulty in implementing computations in a high-through put manner. For example, the COVID-19 Moonshot initiative ran over 5000 free energy simulations exploiting global Folding@home computational initiative (32). This massive undertaking used hundreds of millions of computer hours to achieve 100-fold improvement in potency against 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 analogues of a lead compound by coupling such calculations with an active learning approach, which included an automatic machine learning model selection.

Using 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 have 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 lead optimization stages in drug design projects by leveraging modern computational approaches.

Methods

Dataset screening and preparation, molecular docking, molecular dynamics and thermodynamic integration details are described in Supplementary Information.

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. - by grants NSF DMS-1563291, MCB-1818213, 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. W. L. Jorgensen, "Progress and issues for computationally guided lead discovery and optimization" in Drug Design: Structure- and Ligand-Based Approaches, C. H. Reynolds, D. Ringe, J. K. M. Merz, Eds. (Cambridge University Press, Cambridge, 2010), DOI: 10.1017/CBO9780511730412.003, pp. 1-14.
2. T. Steinbrecher, "Free Energy Calculations in Drug Lead Optimization" in Protein-Ligand Interactions. (2012), https://doi.org/10.1002/9783527645947.ch11, pp. 207-236.
3. O. O. Grygorenko *et al.*, Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **23**, 101681 (2020).
4. J. Lyu *et al.*, Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224-229 (2019).
5. A. A. Sadybekov *et al.*, Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **601**, 452-459 (2022).
6. J. G. Kirkwood, Statistical Mechanics of Fluid Mixtures. *J. Chem. Phys.* **3**, 300-313 (1935).
7. A. S. J. S. Mey *et al.*, Best Practices for Alchemical Free Energy Calculations [Article v1.0]. *Living J. Comp. Mol. Sci* **2**, 18378 (2020).
8. L. Wang *et al.*, 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.* **137**, 2695-2703 (2015).

9. L. F. Song, T.-S. Lee, C. Zhu, D. M. York, K. M. Merz, Using AMBER18 for Relative Free Energy Calculations. *J. Chem. Inf. Model.* **59**, 3128-3135 (2019).
10. M. Kuhn *et al.*, Assessment of Binding Affinity via Alchemical Free-Energy Calculations. *J. Chem. Inf. Model.* **60**, 3120-3130 (2020).
11. X. He *et al.*, 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* **5**, 4611-4619 (2020).
12. K. D. Konze *et al.*, 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.* **59**, 3782-3793 (2019).
13. R. Abel, L. Wang, L. D. Mobley, A. R. Friesner, A Critical Review of Validation, Blind Testing, and Real- World Use of Alchemical Protein-Ligand Binding Free Energy Calculations. *Current Topics in Medicinal Chemistry* **17**, 2577-2585 (2017).
14. D. Shin *et al.*, Papain-like protease regulates SARS-CoV-2 viral spread and innate immunity. *Nature* **587**, 657-662 (2020).
15. J. Osipiuk *et al.*, Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. *Nat. Commun.* **12**, 743 (2021).
16. Z. Shen *et al.*, Design of SARS-CoV-2 PLpro Inhibitors for COVID-19 Antiviral Therapy Leveraging Binding Cooperativity. *J. Med. Chem.* **65**, 2940-2955 (2022).
17. B. Settles, Active learning literature survey. (2009).
18. C. Ambroise, G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 6562 (2002).
19. T. Ahneman Derek, G. Estrada Jesús, S. Lin, D. Dreher Spencer, G. Doyle Abigail, Predicting reaction performance in C–N cross-coupling using machine learning. *Science* **360**, 186-190 (2018).
20. V. Chuang Kangway, J. Keiser Michael, Comment on "Predicting reaction performance in C–N cross-coupling using machine learning". *Science* **362**, eaat8603 (2018).
21. P. M. R. DeVries, F. Viégas, M. Wattenberg, B. J. Meade, Deep learning of aftershock patterns following large earthquakes. *Nature* **560**, 632-634 (2018).
22. A. Mignan, M. Broccardo, One neuron versus deep learning in aftershock prediction. *Nature* **574**, E1-E3 (2019).
23. X. He, K. Zhao, X. Chu, AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems* **212**, 106622 (2021).
24. L. Tuggener *et al.* (2019) Automated Machine Learning in Practice: State of the Art and Recent Results. in *2019 6th Swiss Conference on Data Science (SDS)*, pp 31-36.
25. K. Chauhan *et al.* (2020) Automated Machine Learning: The New Wave of Machine Learning. in *2020 2nd International Conference on Innovative Mechanisms for Industry Applications (ICIMIA)*, pp 205-212.
26. M. Feurer *et al.*, Efficient and robust automated machine learning. *Advances in neural information processing systems* **28** (2015).
27. M. Feurer *et al.*, "Auto-sklearn: Efficient and Robust Automated Machine Learning" in Automated Machine Learning: Methods, Systems, Challenges, F. Hutter, L. Kotthoff, J. Vanschoren, Eds. (Springer International Publishing, Cham, 2019), 10.1007/978-3-030-05318-5_6, pp. 113-134.
28. F. Pedregosa *et al.*, Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825-2830 (2011).
29. D.A. Case, I.Y. Ben-Shalom, S.R. Brozell, D.S. Cerutti, T.E. Cheatham, III, V.W.D. Cruzeiro, T.A. Darden, R.E. Duke, D. Ghoreishi, M.K. Gilson, H. Gohlke, A.W. Goetz, D. Greene, R Harris, N. Homeyer, Y. Huang, S. Izadi, A. Kovalenko, T. Kurtzman, T.S. Lee, S. LeGrand, P. Li, C. Lin, J. Liu, T. Luchko, R. Luo, D.J. Mermelstein, K.M. Merz, Y. Miao, G. Monard, C. Nguyen, H. Nguyen, I. Omelyan, A. Onufriev, F. Pan, R. Qi, D.R. Roe, A. Roitberg, C. Sagui, S. Schott-Verdugo, J. Shen, C.L. Simmerling, J. Smith, R. SalomonFerrer, J. Swails, R.C. Walker, J. Wang, H. Wei, R.M. Wolf, X. Wu, L. Xiao, D.M. York and P.A. Kollman (2018), AMBER 2018, University of California, San Francisco.
30. H. B. Mann, D. R. Whitney, On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics* **18**, 50-60, 11 (1947).
31. L. Van der Maaten, G. Hinton, Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9** (2008).
32. F. von Delft *et al.*, A white-knuckle ride of open COVID drug discovery. *Nature* **594**, 330-332 (2021).