Title: Interpretable deep-learning pK_a prediction for small molecule drugs via atomic sensitivity analysis

Authors: Joseph DeCorte[1-3], Benjamin Brown[4,5], Jens Meiler[1,2,4-6*]

Affiliations:

[1]Department of Chemical and Physical Biology, Vanderbilt University, Nashville, TN 37232, USA.

[2]Center for Structural Biology, Vanderbilt University, Nashville, TN 37232, USA.

[3]Vanderbilt Medical Scientist Training Program Vanderbilt University Medical Center, Vanderbilt University School of Medicine, Nashville TN, 37232-8725, USA.

[4]Department of Chemistry, Vanderbilt University, Nashville, TN 37232-8275, USA.

[5]Center for Applied AI in Protein Dynamics, Vanderbilt University, Nashville, TN 37232-8725, USA

[6]Institute for Drug Discovery, Leipzig University Medical School, Leipzig, SAC 04103, Germany

*Corresponding author. Email: jens.meiler@vanderbilt.edu

ORCID

Joseph DeCorte: orcid.org/0000-0002-0131-6176

Benjamin Brown: *orcid.org/0000-0001-5296-087X*

Jens Meiler: *orcid.org/0000-0001-8945-193X*

ABSTRACT

Machine learning (ML) models play a crucial role in predicting properties essential to drug development, such as a drug's logscale acid-dissociation constant (pK_a). Despite recent architectural advances, these models often generalize poorly to novel compounds due to a scarcity of ground-truth data. Further, these models lack interpretability, in part due to a dependence on explicit encodings of input molecules' molecular substructures. To this end, atomic-resolution information is accessible in chemical structures by observing model response to atomic perturbations of an input molecule; however, no methods exist that systematically utilize this information for model and molecular analysis. Here, we present BCL-XpKa, a substructure-independent, deep neural network (DNN)-based pK_a predictor that generalizes well to novel small molecules. BCL-XpKa discretizes pK_a prediction from a regression problem into a multitask-classification problem, which accumulates data for prediction at biologically relevant pH values and records the model's uncertainty in its prediction as a discrete distribution for each

pK_a prediction. BCL-XpKa outperforms modern ML pK_a predictors and accurately models the effects of common molecular modifications on a molecule's ionizability. We then leverage BCL-XpKa's substructure independence to introduce atomic sensitivity analysis (ASA), which quickly decomposes a molecule's predicted pK_a value into its respective atomic contributions without model retraining. When paired with BCL-XpKa, ASA informs that BCL-XpKa has implicitly learned high-resolution information about molecular substructures. We further demonstrate ASA's utility in structure preparation for protein-ligand docking by identifying ionization sites in 97.8% and 83.4% of complex small molecule acids and bases. We then apply ASA with BCL-XpKa to understand the physicochemical liabilities and guide optimization of a recently published KRAS-degrading PROTAC.

INTRODUCTION

Predicting a drug's behavior in the body is a key challenge in computational drug development. For example, accurate prediction of compounds' bioavailability could support early modification or termination of nonviable lead molecules, thereby saving years of time and millions of dollars on research and development. The demand for fast and accurate predictions of a drug's quantitative structure-activity and structure-property relationships (QSAR, QSPR) has skyrocketed as our access to synthesizable chemical space approaches one trillion molecules[1]. While advances in machine learning have improved prediction accuracy, the small amount of publicly available, high-quality experimental data for training often leads to overfitting and prevents generalizability[2-4]. Further, QSPR model interpretability is often poor despite the relatively more intuitive input (chemical structures) than in other fields of computational biology (e.g., transcriptomics data). As such, additional explorations into architectures that can efficiently train on chemical data, as well as general methods to interpret these models' outputs, are warranted.

One of the most critical properties to a drug's downstream efficacy is its ionizability at physiologic pH values, which depends on the drug's logscale acid-dissociation constant (pK_a value)[4,5]. Quantum mechanical (QM) methods now calculate pK_a with experimental accuracy and are extremely valuable to late-stage drug development. However, small-molecule drug development often begins with virtual high-throughput screening (vHTS) of billions of compounds, and QM methods are too computationally expensive to assist meaningfully in vHTS. As such, scientists have made tremendous investment in ML-based QSAR/QSPR predictors for faster – though potentially less accurate – prediction of physicochemical properties like pK_a.

These ML methods generally embed molecules using molecular fingerprints, which are two-dimensional (2D)- or 3D chemical substructures centered around each atom in the molecule. Recently, groups have realized significant gains in prediction accuracy with graph neural networks (GNNs), which embed molecules as a graph in addition to standard chemical descriptors[6-8]. Improvements in molecular featurization strategies and network architecture have driven state-of-the-art pK_a prediction accuracy to within 0.75-1.00 pK_a units of experimental values.

Despite these advances, several limitations persist in ML-based pK_a prediction and QSPR prediction generally. First, all ML-based pK_a predictors to date use regression. While regression is the natural setting for predicting continuous values, small training set sizes restrict the accuracy of regression outputs, particularly at extreme – but still physically relevant – values. Second, many of these models directly encode common molecular substructures in their feature set[9]. This strategy may limit generalizability by preventing complete consideration of each atom's local context in a molecule. For example, amide Nitrogen atoms are generally not ionizable at physiologic pH values, but their acidity can be greatly increased by appending neighboring electron-withdrawing groups (e.g., diacetamide). Therefore, encoding atom-specific local environments may increase generalizability.

Finally, model explainability and interpretability in computational chemistry largely focus on feature-set-level analysis, encompassing step-wise[10,11], feature-masking[12], feature-set-perturbation[13], feature-attribution[14], and response-randomization[15] methods. However, feature-set-level analysis is often slow and unintuitive, as feature sets are often large and complex. Recently, ML frameworks have been developed that utilize direct chemical representations (as SMILES strings) to identify important substructures for transformer predictions[16], but no group to date has leveraged perturbations to the input chemical structure itself to gain insights into model learning and output. Indeed, chemical structures are unique in computational biology in that they can be perturbed in consistent, physical meaningful ways. With an appropriately constructed feature set, measuring a model's response to these perturbations would provide granular details into both model learning and molecular hotspots for prediction in real time, without the need for model retraining. For example, replacing a molecule's acidic carboxylic acid functional group with an inert ketone increases the pK_a from ~4 to ~20, thereby demonstrating the carboxylic acid's importance to acidity. As a counterexample, increasing or decreasing the expression of a gene in a transcriptomics-based predictor of cellular activity may not be physically meaningful, as gene expression is highly dependent on the network of expressed genes in a cell/tissue. This presents an underutilized opportunity for computational chemists to gain valuable insights into model learning, performance, and molecular structure.

To address these limitations, we present BCL-XpKa, a substructure-independent multitask classifier for rapid and accurate pK_a prediction built in the Biology and Chemistry Library (BCL), an open-source cheminformatics platform developed and maintained by our lab. We use pK_a prediction to illustrate that discretizing continuous problems in chemical biology into multitask classification problems can increase prediction accuracy without meaningful information loss, thereby circumventing many of the problems associated with regression models. We couple BCL-pKa with a novel method of atomic sensitivity analysis (ASA), which provides unprecedented, atomic-level insights into which regions of a molecule are most important for the model's final prediction. We demonstrate ASA's utility in probing model learning, as well as its direct applicability to introducing targeted modifications in molecules that reduce ionizability. Importantly, ASA is relevant to all forms of QSAR/QSPR prediction and can be easily implemented to existing algorithms.

BCL-XpKa is a multi-layer perceptron (MLP) that embeds molecules using 2D chemical descriptors features that only encode information about each atom's local environment (up to 1 bond away) in the molecule[17]. This scheme is substructure independent and enables increased sensitivity to atom-level perturbations, which is particularly important for hit-to-lead and lead optimization in late-stage drug development. Small molecule drugs often have both basic and acidic regions that vary greatly in pK_a values (e.g., amino acids have both an acidic carboxylic acid and a basic free amine group). To account for this, we trained two models: one to predict a molecule's most acidic pK_a value (BCL-XpKaAcid), and one to predict its most basic pK_a value (BCL-XpKaBase). This is a common practice in modern pK_a prediction. We trained BCL-XpKa on datasets of both predicted and experimental pK_a values, and we evaluate our models on an external test set of challenging acids and bases with experimental pK_a values. We find that our model has competitive accuracy and reduced substructure dependence than state-of-the-art pK_a predictors, including GNN-based models.

Overall, the work presented here has the following significant contributions to the field:

- We developed a novel, substructure-independent framework for QSPR prediction that uses local atomic environment embeddings and replaces regression with multitask classification, using pK_a prediction to illustrate competitive performance with modern ML models.
- We developed a method that rapidly assesses QSPR model learning and provides atomic-level insights to molecular ionizability without requiring model retraining
- We integrate these two tools in a workflow for lead optimization and apply it to optimize a pan-KRAS degrading Proteolysis Targeting Chimera (PROTAC).

RESULTS

Multitask Classification is competitive with Regression for small-molecule pK_a prediction

Regression models naturally dominate machine-learning (ML) approaches to QSPR prediction. However, regression models require large, high-quality datasets to train properly, and understanding the model's uncertainty in its prediction is challenging. Here, we construct BCL-XpKa, a multilayer perceptron (MLP)-based pK_a prediction tool that transforms a continuous prediction problem into a multitask classification problem. Rather than predict pK_a values on a continuous range, BCL-XpKa predicts the probability that a molecule's pK_a lies within a certain range (Figure 1A-B). The expected value of this probability distribution corresponds to BCL-XpKa's predicted pK_a for a molecule, and the variance of this distribution directly informs its confidence in that prediction.

To train BCL-XpKa, pK_a values in the training set were converted into vectors in $\mathbb{Z}_2^n$, where n is the number of bins the continuous interval has been divided into, and a 1 at position *i* indicates the molecule's pK_a lies in bin *i* (Figure 1B). BCL-XpKa models were trained to predict the most acidic and most basic pK_a values of a molecule using training data from ChEMBL augmented with negative data (i.e., nonionizable molecules).

Binning training data in this way necessarily leads to some information loss. Using an external test set, we demonstrate that mean absolute error (MAE) increases as the number of bins increases, as increasing the number of bins reduces the amount of data in each bin for training (Figure 1C). A bin size of 1 pK_a unit yielded the lowest MAE on this test set and is used in the BCL-XpKa production model. For both acids and bases, BCL-XpKa marginally outperforms the best-performing regression models trained on the same data (0.79 vs 0.83 for acids, 0.86 vs 0.92 for bases, Figure 1D). Additional model details, including an evaluation of model hyperparameters, can be found in Supplemental Figure 1.

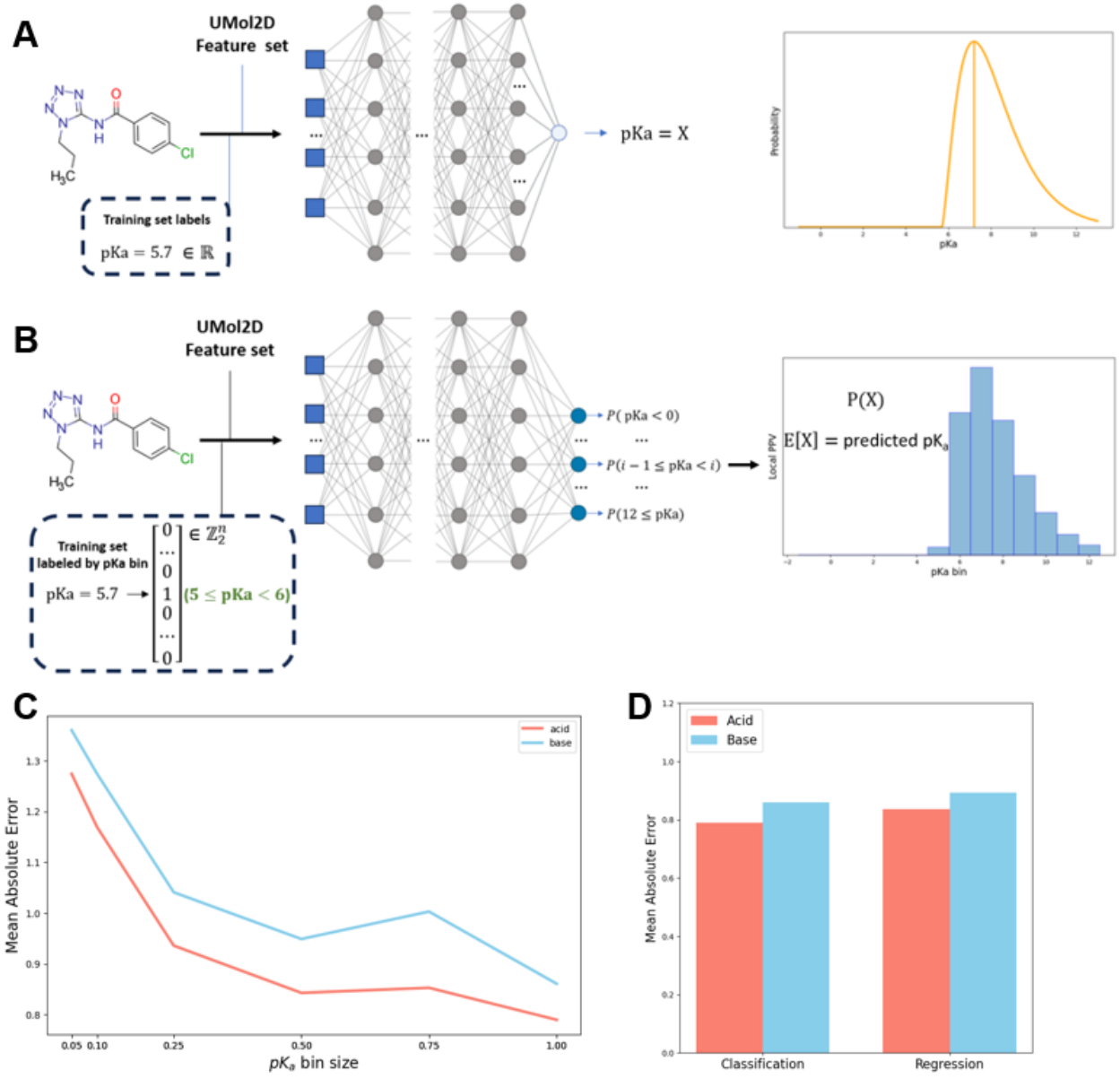

Figure 1 BCL-XpKa model description and internal performance (A-B) Overview of the regression and MTC architectures for pK_a prediction. BCL-XpKa utilizes the MTC architecture with bin size of 1 pK_a unit. (C) MTC model performance by pK_a bin size on acids (red) and

bases (blue). (D) Model performance comparison between best performing MTC (BCL-XpKa) and regression architectures on acids (red) and bases (blue).

BCL-XpKa accurately captures complex trends in ionizability for druglike small molecules

We then compared BCL-XpKa's performance head-to-head against several state-of-the-art pK_a predictors of varying architecture using a challenging test set external to BCL-XpKa's training data[7,18-21]. Despite its relatively simple architecture, BCL-XpKa achieves competitive performance to the best machine-learning (MolGpKa, graph-convolutional neural network) and rule-based pK_a predictors on both acids and bases, with BCL-XpKaAcid and BCL-XpKaBase achieving mean absolute error (MAE) of 0.79 and 0.86, respectively (Figure 2A).

Beyond accurate pK_a prediction, correctly predicting the effect of small perturbations to a lead molecule's ionizability is of key importance in drug development[22]. To assess BCL-XpKa's sensitivity to such changes, we identified 71 pairs of molecules in our test set that vary by a slight modification, such as the replacement of an amide with an ester. BCL-XpKa correctly predicts the direction of pKa change in 81.7% of these pairs (Figure 2B). To illustrate this effect, BCL-XpKa correctly predicts the inductive effect of electron-withdrawing groups on acidity using a series of phenol derivatives (Figure 2C). Here, fluorination at the ortho position increases acidity more than fluorination at the para position (8.74 vs 9.28), and substitution with multiple fluorine atoms has a greater effect than monosubstitution (7.36). While phenol was in our training data, the remaining molecules were not. For bases, BCL-XpKa correctly predicts the complex impact of aromaticity on nitrogen basicity in a series of piperidine derivatives relevant to drug development. Introducing a neighboring phenyl group reduces predicted pK_a from 10.45 (true pK_a 11.2) to 4.80 (5.00). Similarly, aromatization of piperidine to pyridine decreases predicted pK_a to 5.45 (5.20), and appending the same phenyl group to produce quinoline reduces pK_a to 4.54 (4.92) (Figure 2D).

Finally, substructure independence is critical to QSPR model generalizability to novel compounds. As described above, BCL-XpKa embeds molecules solely using the 1-bond-length neighborhoods of each atom to limit substructure dependence. To investigate this strategy's impact, we subset BCL-XpKa's training set according to 30 ionizable functional groups. We iteratively retrained BCL-XpKa leaving each substructural class out, then tested each model on its withheld substructural class (Figure 2E). BCL-XpKa demonstrates robust performance on this leave-class-out (LCO) test, with an average MAE of 1.1 pK_a units across all LCO models. Training on MACCS descriptors rather than Mol2D yielded systematically worse results and an average MAE of 1.46 pK_a units (Figure 2F).

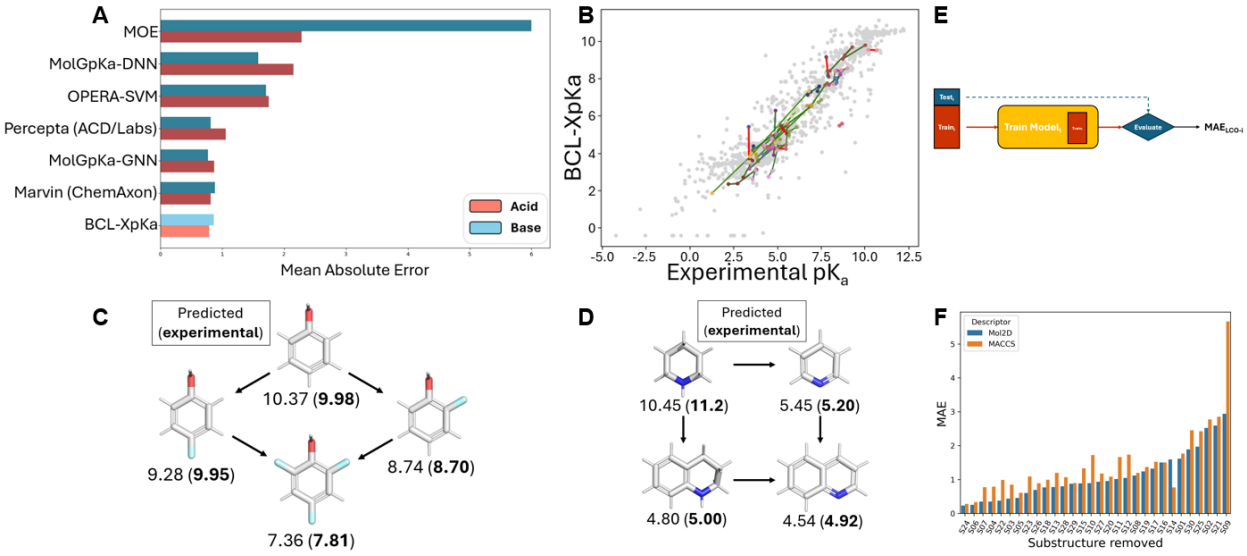

Figure 2: BCL-XpKa external performance and molecular series (A) Performance of various pK_a predictors on an external test set of acids (red) and bases (blue). (B) BCL-XpKa prediction vs experimental pK_a value for families of related druglike molecules. Green denotes correct change in predicted pK_a due to chemical modification, and red denotes incorrect change in predicted pK_a. (C-D) Example molecular families from (B). Predicted and experimental pK_a values provided. (E) Schematic for LCO testing. $Test_i$ and $Train_i$ denote the subsets of the training set that contain or do not contain, respectively, substructure i. $Model_i$ was trained with $Train_i$ and evaluated on $Test_i$ to give $MAE_{LCO\text{-}i}$. (F) LCO performance of BCL-XpKa (blue) vs an equivalent model trained with a MACCS-based descriptor set (orange).

Atomic sensitivity analysis provides actionable, atomic-resolution information on model predictions

Computational chemistry currently lacks rapid, reliable tools for interpreting ML model predictions[23]. Such atom-, substructure-, or pharmacophore-level information could accelerate computer-aided drug development on multiple fronts, from assisting in model training, to preparing and filtering molecules in virtual high-throughput screens, to guiding lead compound modification in lead optimization. To address this deficit, we developed atomic sensitivity analysis (ASA, Figure 3A), and we demonstrate its utility in decomposing BCL-XpKa's predictions to assess model learning, identify ionization sites in complex small molecules, and guide lead optimization efforts by reducing molecular ionizability.

ASA compares an ML model's prediction on a parent molecule before and after some perturbation. Here, we sequentially replace heteroatoms in the parent molecule with correctly hybridized carbons, then generate probability distributions for each with BCL-XpKa. The final ASA score is a scaled version of the KL-Divergence between these parent and perturbed pKa distributions (see *Methods*). Importantly, we anticipate that, with careful feature-set selection,

this replace-rescore-compare scheme will generalize well to substructure- and pharmacophore-level perturbations (Figure 3A).

We benchmarked ASA on BCL-XpKa on all test-set molecules with nontrivial ionization sites. We first hypothesized that perturbing an acid's most acidic hydroxyl group or a base's most basic Nitrogen atom would have the most significant impact on the predicted pKa, and therefore the largest ASA scores. We tested this hypothesis by performing ASA on the Oxygen atoms in the acid set and the Nitrogen atoms in the base set and considering only the atom with the maximum ASA score in each molecule. This strategy correctly identifies 97.8% of the most acidic Oxygen atoms in the acid set and 83.4% of the most basic Nitrogen atoms in the base set, thereby demonstrating ASA's potential utility in high-throughput structure preparation (Figure 3B).

This benchmark revealed surprisingly consistent ASA scores for each atom in the substructures that recurred throughout the test set. For example, free amines are the most basic group in 33.9% of our experimentally characterized bases, and in each of these molecules the amine Nitrogen atom dominates the molecule's ASA scores (33.1 +/- 16.0). These scores were significantly higher than average scores for Nitrogen atoms in amide (0.225 +/- 0.332), indole (0.261 +/- 0.444), and nitrile groups (0.180 +/- 0.357) ($p < 0.001$), functional groups which are not traditionally ionizable at physiologic pH and which dominated 0% of the test-set ASA scores (Figure 3C). Further, molecules where the dominant ionizable group is less basic than typical amines also demonstrated consistent ASA scores, and this ASA dominance was reliably ablated by the introduction of an amine functional group (Figure 3D-E).

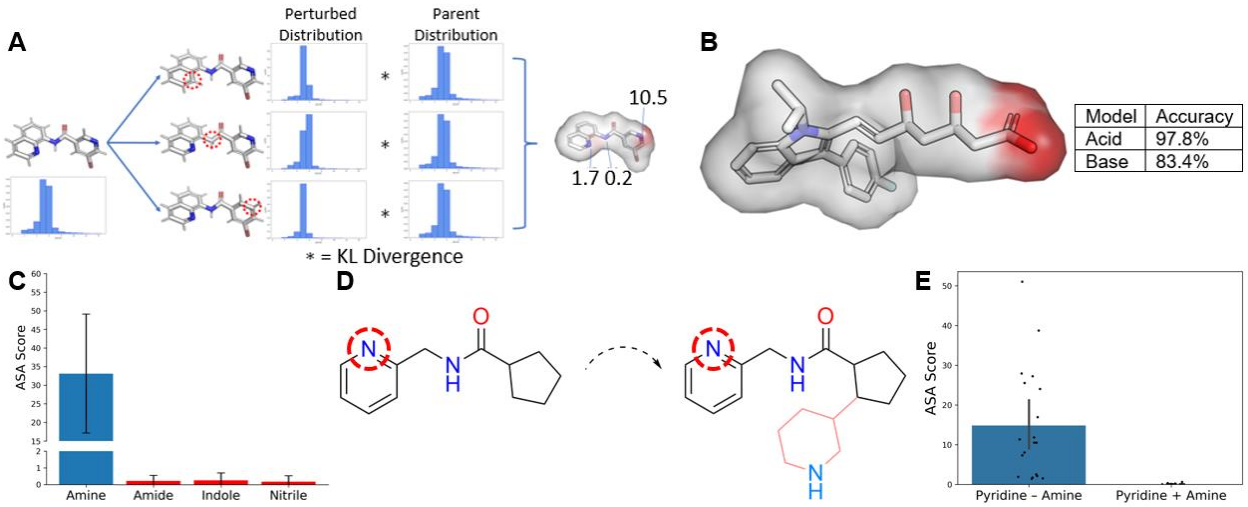

Figure 3: Atomic sensitivity for molecular analysis (A) Schematic of the ASA protocol. Parent and Perturbed distributions refer to the localPPV distributions output by BCL-XpKa. (B) ASA accuracy at detecting the most acidic Oxygen atom and most basic Nitrogen atom in all non-trivial test-set molecules. (C) ASA scores of positive- and negative-control substructures for BCL-XpKaBase decomposition. Blue denotes the positive control, red denotes the negative controls. (D-E) Modulation of pyridine Nitrogen ASA score by addition of an amine group.

ASA reveals BCL-XpKa implicitly learns substructural information despite substructure-free embeddings

To investigate ASA scoring consistency further, we performed ASA on Nitrogen atoms in the most frequently occurring substructures in the basic test set, beyond the controls discussed above. All subgroups examined demonstrated statistically significant loss of ASA signal when a more dominant subgroup (i.e., more relevant to the molecule's basicity) was present (Figure 4A-D), indicating that retaining the most dominant ASA atom also preserves BCL-XpKa's predicted pK_a distribution for the molecule.

Filtering out these non-dominant molecules reveals that most substructures have consistent, substructure-specific trends in ASA scores (Figure 4E). This substructure-specificity even persists when separate substructures have Nitrogen atoms with identical local atomic environments. For example, indole and imidazole both have a Nitrogen bound to a Hydrogen and two sp^2 Carbon atoms, but only imidazole is ionizable at physiologic pH values[24,25]. While ASA correctly distinguishes that imidazole's other Nitrogen (which has a lone pair of electrons) is the most basic Nitrogen in imidazole, it also scores the N-H Nitrogen significantly higher than the identical motif in indole (Figure 3C), suggesting that this Nitrogen is critical to imidazole's observed basicity.

Some of these structures have surprisingly high variance in ASA scores, particularly the pyridine and aniline substructures. From manual inspection, we hypothesized a portion of this variance is attributable to the impact of neighboring electron-donating and -withdrawing groups (EDGs, EWGs), which respectively increase and decrease basicity of neighboring Nitrogen atoms. We tested this hypothesis by scoring manually created sets of pyridine derivatives with various EDG and EWG substituents, which confirmed as suspected that neighboring EDGs tend to increase ASA scores, and neighboring EWGs tend to decrease ASA scores (Figure 4F). Interestingly, symmetric substructures also contributed to this variance, as the symmetric substructure masks the effect of the removed atom during ASA scoring (Figure 4G).

Together, these ASA findings suggest that BCL-XpKa has learned impressive substructural insights that are adaptable to molecular context without directly encoding these substructures in the feature set.

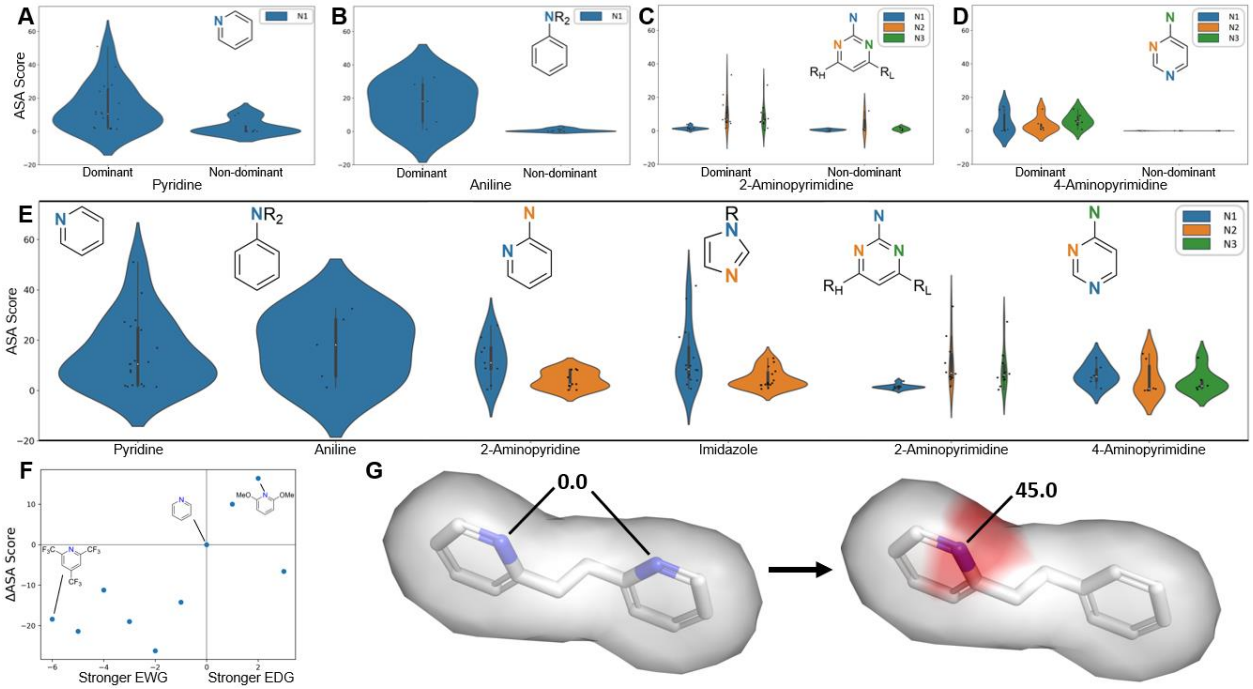

Figure 4: Atomic sensitivity analysis of substructures (A-D) Violin plots of ASA scores for commonly occurring substructures when these substructures are the dominant site of a molecule's ionization vs when a more dominant substructure was present. (E) Violin plots of ASA scores for commonly occurring substructures when these substructures were the dominant site of ionization. Notably, all test-set bases containing 2-Aminopyridine and Imidazole featured them as their dominant ionization site. (F) Change in pyridine Nitrogen's ASA Score by neighboring EWG or EDG groups. (F) Masking effect of molecular symmetry on ASA score. ASA = Atomic Sensitivity Analysis; EWG = electron-withdrawing group; EDG = electron-donating group.

Atomic sensitivity analysis can inform lead compound optimization in drug development

Atomic sensitivity analysis also has promising utility in prospective drug design. Significant interest has developed in the past two decades in targeted protein degradation via small-molecule Proteolysis Targeting Chimeras (PROTACs). PROTACs consist of two small molecules joined by a linker and form flexible ternary complexes with the target and an E3 Ligase, which allows for target ubiquitination and subsequent degradation by the 26S proteasome (Figure 5A, PDB: 8QU8). PROTACs degrade targets catalytically, making them an attractive strategy for challenging targets that have evaded small-molecule inhibition; however, PROTAC size and complexity plagues design efforts with poor bioavailability and cell permeability[26]. These properties are generally optimized through modification of the PROTAC linker, as the Ligase- and target-binding domains make specific contacts with their respective proteins. Here, we demonstrate how atomic sensitivities can guide rational changes to the PROTAC linker to minimize PROTAC ionizability.

KRAS is a challenging target in oncology because it lacks a deep, well-defined pocket for small-molecule inhibition[27]. Recently, Popow et al. created a KRAS-degrading PROTAC using the VHL E3 Ligase (Figure 5A-B)[28]. Our model predicts this PROTAC has a pK_a of 6.51 (**P-1**, Figure 4B) , suggesting protonatability at physiologic pH values. Atomic sensitivity analysis of the linker reveals one of two tertiary amines in P-1 drives this prediction (Figure 5C). The crystal structure of the P-1 ternary complex demonstrates this amine forms a salt bridge with KRAS Q62 when protonated. While salt-bridge interactions promote strong drug binding, PROTACs only need to bind their targets transiently, and a protonatable amine is a liability for permeability. Based on this analysis, we evaluated several P-1 bioisosteric modifications at this amine that reduce predicted pK_a (Figure 5D).

We evaluated each modification's performance computationally using a well-benchmarked ternary complex (TC) generation algorithm, which produces a TC ensemble and "double clusters" this ensemble by structural similarity of both the Ligase-Target interactions and PROTAC conformation[29]. Per the algorithm authors' benchmark, the largest structural double cluster often yields the most crystal-structure-like poses and filters out false positives from the full ensemble. Each modification produced comparable TC ensembles and double clusters to P-1 (Figure 5E).

Emerging experimental evidence suggests that a PROTAC's degradation efficiency depends on its ability to place the target protein in proximity to ubiquitin (<60Å) in the complete E3 Ligase complex[30]. As such, we also evaluated the modifications' ability to place KRAS in proximity to ubiquitin by superposing each TC ensemble member into the full E3 Ligase complex. Here, we find that the amide modification in P-2 provides KRAS similar ubiquitin access to P-1 and superior access to all other modifications tested (Figure 5F). Furthermore, docking P-2 into the VHL-KRAS pose identified in the 8QU8 crystal structure demonstrates that P-2 can recapitulate the binding pose of P-1 (heavy atom RMSD <2.5Å), including the key hydrogen bond with KRAS Q62 that P-1 utilizes using the tertiary amine (Figure 5G).

Together, these results demonstrate a computationally validated use of atomic sensitivities to guide lead-molecule optimization.

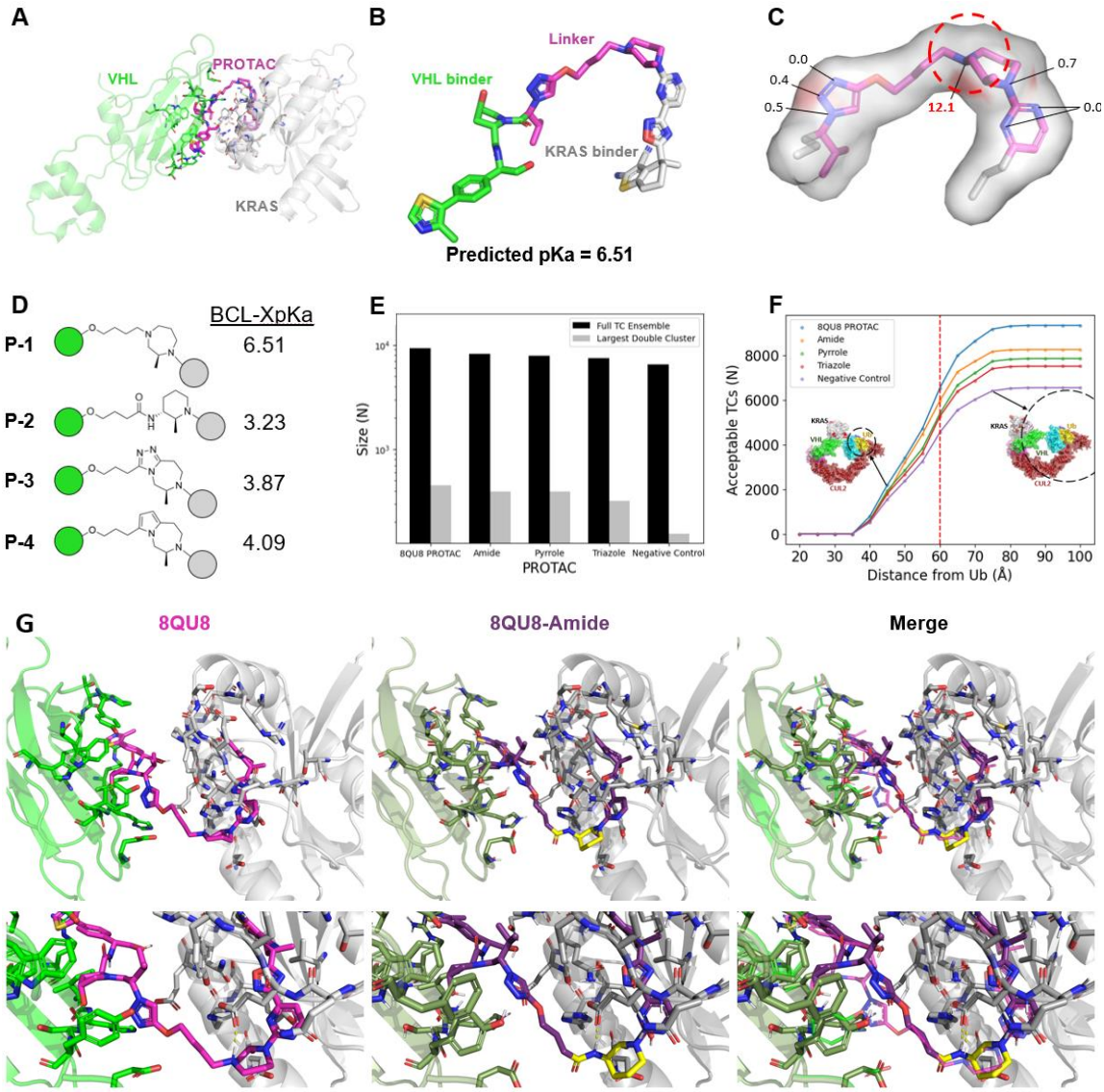

Figure 5: Atomic sensitivity for drug design (A) Crystal structure (PDB: 8QU8) of pan-KRAS degrading PROTAC **P-1** in ternary complex with VHL and KRAS. (B) PROTAC **P-1** colored according to 5A, with pK_a calculated by BCL-XpKa. (C) ASA scores for **P-1** Nitrogen atoms. VHL- and KRAS-binders omitted for space. (D) Proposed bioisosteric **P-1** linker modifications with pK_a values predicted by BCL-XpKa. (E) Ternary-complex ensemble size and size of largest Protein-PROTAC conformational double cluster for each linker modification in 5D. (F) Cumulative number of ternary complexes near Ubiquitin at increasing distances from Ubiquitin in the closed conformation of the CUL2 E3 Ligase complex. Red line at 60Å denotes an empirically estimated distance beyond which target ubiquitination is improbable. (G) Representative images of the 8QU8 crystal structure and the Amide-based linker modification **P-2** supporting similar PROTAC conformations that preserve the hydrogen bond to KRAS Q62. P-1 complex shown in brighter colors; P-2 complex shown in muted colors; P-2 linker modification highlighted in yellow.

DISCUSSION

Here, we have presented BCL-XpKa, a deep-learning based pK_a predictor that reframes QSPR prediction as a classification problem and avoids explicit substructural embeddings while maintaining competitiveness with contemporary machine learning pK_a predictors. We found that this multitask classification approach directly informs the model's uncertainty in its prediction, and that, beyond its absolute accuracy, BCL-XpKa reliably predicts the effects of common molecular modifications made to a hit/lead compound in a drug development program. We also showed that BCL-XpKa generalizes to foreign substructures better than equivalent models trained on MACCS-based descriptors via leave-substructural-class-out validation.

We then used BCL-XpKa as a model system to introduce atomic sensitivity analysis (ASA), a first-in-class ML interpretability method we designed to provide actionable insights into QSPR model output by decomposing a molecule's QSPR prediction into its atomic contributions through direct perturbation of the input chemical structure. When applied to BCL-XpKa, ASA identifies the most ionizable atoms in both acids and bases with remarkable accuracy. ASA also revealed surprisingly consistent results for how BCL-XpKa considers ionizable substructures at the atomic level. These substructural ASA scores were responsive to neighboring electron donating and withdrawing groups, demonstrating that BCL-XpKa learns context-dependent substructural information without explicit substructural embeddings. Finally, we showed that pairing a QSPR model's molecule-level predictions with atomic-level contributions is a powerful tool for guiding lead optimization using a published KRAS-degrading PROTAC. Here, BCL-XpKa and ASA directed linker modifications that reduced PROTAC ionizability while retaining critical PROTAC-KRAS contacts from the original crystal structure.

Several limitations exist in our current framework. First, while regression models can predict arbitrarily extreme values given enough quality data, BCL-XpKa must place all extreme values in two catch-all bins, "$pK_a < 0$" and "$pK_a > 12$", given its multitask classifier architecture. While this limits BCL-XpKa's theoretical output range to -0.5 to 12.5, this is not consequential for biologically relevant pH scales and only marginally affects prediction accuracy.

Further, ASA is currently limited to atomic-level model explainability. While this provides excellent resolution for atomic properties like pK_a, there are many QSPR tasks where understanding the contribution of entire substructures or pharmacophores would be valuable. Generalizing ASA to higher order molecular substructures will further expand our understanding of QSPR ML model predictions and allow ASA to be tailored to specific tasks.

BCL-XpKa and ASA have fundamental applications in computational chemistry generally, as well as early- and late-stage drug development. First, ASA is a generalizable strategy that can increase the explainability of any machine learning model that uses chemical structures as input data. As shown here, ASA scores can help scientists understand what their model has learned from their training set. This information can then guide training-set data augmentation or feature-set modifications.

Further, BCL-XpKa paired with ASA is positioned well to support high-quality small-molecule structure preparation for virtual high-throughput screening (vHTS). vHTS involves screening

ultra-large libraries (ULLs) of small molecules (currently nearing 10^{11} molecules) for their ability to bind to a protein target. vHTS has notoriously low hit rates, and improper protonation of ULL molecules can contribute to both false-positive and false-negative vHTS screens. BCL-XpKa and ASA's speed and accuracy at predicting pK_a and ionization sites in multiprotic species make this tool a valuable asset for ULL structure preparation and downstream protein-ligand analysis in vHTS.

Finally, as demonstrated here, BCL-XpKa paired with ASA can identify ionizable regions in a compound for modification in hit-to-lead or lead optimization. While ionization-site identification is relatively straightforward, this model-ASA strategy generalizes to any QSPR/QSAR model. Therefore, applying ASA to predictors of ADMET/DMPK may facilitate understanding of important but less readily interpretable liabilities in a hit or lead compound.

METHODS

Training Datasets

ChEMBL27 is an open-source database contains over 2 million molecules with various physicochemical descriptors[31]. ACDlabs was used to calculate acidic pK_a and basic pK_a values (chembl_acid_pka and chembl_base_pka, respectively) for ChEMBL molecules, and molecules that were included in our test sets were excluded[18]. We also generated negative data (molecules with no ionization site) in the BCL and set chembl_acid_pka = 50, chembl_base_pka = 0. In sum, acidic pK_a models were trained on 988,643 molecules, and basic pK_a models on 812,918 molecules.

Molecule Preparation

Molecular 3D structures were standardized using Corina for training and testing BCL-XpKa[32]. For external models, structure preparation followed the authors' direction, and Corina was used if no structure preparation method was mentioned. This standardization was used exclusively for downstream usability, as BCL-XpKa solely uses 2D descriptors. All PROTAC modifications introduced in Figure 5 were minimized in the Molecular Operating Environment (MOE) prior to ternary complex ensemble generation.

Molecular Features

The Mol2D molecular descriptor set was used to encode molecules as described elsewhere[17]. Briefly, for each atom in a molecule, Mol2D encodes information about that atom, the bonds made to that atom, and the atoms one bond length away from that atom.

To train a multitask classifier with N output labels, ChEMBL pK_a values were encoded in an N x 1 result label $R \in \mathbb{Z}_2^N$, where the first entry corresponds to $\mathrm{pK_a} < 0$, the last entry to $\mathrm{pK_a} \geq$

12, and the i^{th} entry to $(12/N) * (i - 1) \leq pK_a < (12/N) * i$. For regression models, the pK_a in the ChEMBL set was used directly as the result label.

Model Training and Validation

Artificial neural networks were built in C++ for the Biology and Chemistry Library (BCL), an open-source cheminformatics platform created and maintained by our lab. Each model was trained for 250 iterations without early stopping, which our lab previously found to be unnecessary when dropout is used[33]. An upper-bound for model performance was calculated through random-split cross validation. A lower-bound for performance was calculated through leave-class-out cross validation (LCO-CV), in which the training set was divided into 30 subset $\{C_i\}_{i=1}^{30}$ based on ionizable groups defined in literature[34]. 30 models were then trained in an iterative all-but-one scheme. Model internal performance was evaluated using the logarithmic receiver operating characteristic curve (logAUC) and AUC using the BCL model:ComputeStatistics application.

Model Output and Evaluation

MTCs with N output labels calculate N local positive predictive values (localPPV), where the i^{th} localPPV denotes the probability that the pK_a lies in the i^{th} pK_a interval (see *Molecular Features* above). For model evaluation, we report mean absolute error (MAE) for reader familiarity. In the supplement, we also provide a Brier score for each model, which is a proper scoring rule[1] for more rigorous evaluation of classification model output.

For each molecule $m_i \in Y$ in a test set Y with M molecules, the pK_a of m_i was encoded with a binary result label $R_i \in \mathbb{Z}_2^N$ as described above. MTC scored each molecule, providing a discrete probability distribution P_i describing the molecule's likely pK_a interval membership. From these distributions, MAE was calculated as:

$$\mathrm{MAE}(\mathrm{P},\mathrm{Y}) = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}}|\mathrm{y_i} - \mathrm{E}[\mathrm{P_i}]|$$

where $E[P_i]$ is the expected value of P_i. Similarly, for several models a Brier score was calculated as:

$$\mathrm{BS}(\mathrm{P},\mathrm{Y}) = \frac{1}{\mathrm{M}}\sum_{\mathrm{i}=1}^{\mathrm{M}}\sum_{\mathrm{j}=1}^{\mathrm{N}}\left(\mathrm{r_{ij}} - \mathrm{p_{ij}}\right)^2$$

[1] "Proper scoring rule" is a term in statistics for a loss function that is minimized if the probability distribution output from the model is identical to the ground-truth probability distribution. When the bidirectional holds, the scoring rule is further labeled a *strictly proper* scoring rule.

Where $r_{ij} \in R_i$ is the j[th] result label for the i[th] molecule in the test set (i.e., whether its pK_a value lies in the j[th] pK_a interval), and p_{ij} is the localPPV that the i[th] molecule's pK_a value lies in the j[th] pK_a interval.

Throughout, MAE is used throughout to compare MTC to Regression and MTC to MTC models. Percent accuracy of categorization is not included, as it is an improper and discontinuous scoring metric.

Atomic sensitivity analysis

Atom replacement schemes were coded in C++ within the BCL. A parent molecule m is scored by BCL-XpKa to produce P, a discrete probability distribution of potential pK_a values. Heteroatom a in the parent molecule is replaced with an appropriately hybridized carbon atom, and the perturbed molecule is rescored to produce P'_a. The dissimilarity between these distributions was calculated by their Kullback-Leibler (KL) divergence:

$$D_{KL}(P'_a||P) = \sum_{j=1}^{N} P'_{a,j} \ln\left(\frac{P'_{a,j}}{P_j}\right),$$

where P_j and $P'_{a,j}$ are localPPVs as described in *Model Output and Evaluation*. Briefly, the KL divergence of these two probability distributions is best interpreted as the relative entropy between these distributions, where $D_{KL}(P'||P) = 0$ denotes the distributions are identical (there would be no "surprises" if a given sample came from P vs P′), and higher values denote more dissimilarity. Finally, KL divergences were empirically denoised to generate ASA scores:

$$ASA(m, a) = e^{\lfloor 5 * D_{KL}(P'_a||P) \rfloor} - 1$$

PROTAC ternary complex ensemble generation

Ternary complexes (TCs) were constructed according to Drummond et al (2020). Briefly, protein-protein interactions (PPIs) with the PROTAC binding pockets near each other, as well as a set of up to 10000 PROTAC conformations, were produced in the Molecular Operating Environment (MOE). PROTAC conformations were then docked into the PPIs and filtered according to the authors' criteria. TCs were then clustered on both protein- and PROTAC-conformations to produce "double clusters." Protein-conformational clustering was done at CA-RMSD < 10Å. PROTAC clustering was done at heavy-atom RMSD < 2.5Å.

Hardware

All models were trained with 18 Intel Xenon W-2295 CPU cores. PROTAC TC formation was performed using an Nvidia RTX A5000 GPU.

REFERENCES

1 Gorgulla, C. Recent Developments in Ultralarge and Structure-Based Virtual Screening Approaches. *Annu Rev Biomed Data Sci* **6**, 229-258, doi:10.1146/annurev-biodatasci-020222-025013 (2023).

2 Dara, S., Dhamercherla, S., Jadav, S. S., Babu, C. M. & Ahsan, M. J. Machine Learning in Drug Discovery: A Review. *Artif Intell Rev* **55**, 1947-1999, doi:10.1007/s10462-021-10058-4 (2022).

3 Maltarollo, V. G., Kronenberger, T., Wrenger, C. & Honorio, K. M. Current Trends in Quantitative Structure–Activity Relationship Validation and Applications On Drug Discovery. *Future Science OA* **3**, FSO214, doi:10.4155/fsoa-2017-0052 (2017).

4 Wu, J., Kang, Y., Pan, P. & Hou, T. Machine learning methods for pKa prediction of small molecules: Advances and challenges. *Drug Discovery Today* **27**, 103372, doi:https://doi.org/10.1016/j.drudis.2022.103372 (2022).

5 Navo, C. D. & Jiménez-Osés, G. Computer Prediction of pKa Values in Small Molecules and Proteins. *ACS Medicinal Chemistry Letters* **12**, 1624-1628, doi:10.1021/acsmedchemlett.1c00435 (2021).

6 Johnston, R. C. *et al.* Epik: pKa and Protonation State Prediction through Machine Learning. *Journal of Chemical Theory and Computation* **19**, 2380-2388, doi:10.1021/acs.jctc.3c00044 (2023).

7 Pan, X., Wang, H., Li, C., Zhang, J. Z. H. & Ji, C. MolGpka: A Web Server for Small Molecule pK(a) Prediction Using a Graph-Convolutional Neural Network. *J Chem Inf Model* **61**, 3159-3165, doi:10.1021/acs.jcim.1c00075 (2021).

8 Xiong, J. *et al.* Multi-instance learning of graph neural networks for aqueous pKa prediction. *Bioinformatics* **38**, 792-798, doi:10.1093/bioinformatics/btab714 (2021).

9 Gierlich, C. & Palkovits, S. Featurizing chemistry for machine learning — methods and a coded example. *Current Opinion in Chemical Engineering* **37**, 100840, doi:https://doi.org/10.1016/j.coche.2022.100840 (2022).

10 BALLS, G. R., PALMER-BROWN, D. & SANDERS, G. E. Investigating microclimatic influences on ozone injury in clover (Trifolium subterraneum) using artificial neural networks. *New Phytologist* **132**, 271-280, doi:https://doi.org/10.1111/j.1469-8137.1996.tb01846.x (1996).

11 Maier, H. R. & Dandy, G. C. The Use of Artificial Neural Networks for the Prediction of Water Quality Parameters. *Water Resources Research* **32**, 1013-1022, doi:https://doi.org/10.1029/96WR03529 (1996).

12 Štrumbelj, E., Kononenko, I. & Robnik Šikonja, M. Explaining instance classifications with interactions of subsets of feature values. *Data & Knowledge Engineering* **68**, 886-904, doi:https://doi.org/10.1016/j.datak.2009.01.004 (2009).

13 Fong, R. C. & Vedaldi, A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3449-3457 (2017).

14 McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proc Natl Acad Sci U S A* **116**, 11624-11629, doi:10.1073/pnas.1820657116 (2019).

15 Olden, J. D. & Jackson, D. A. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. *Ecological Modelling* **154**, 135-150, doi:https://doi.org/10.1016/S0304-3800(02)00064-9 (2002).

16 Karpov, P., Godin, G. & Tetko, I. V. Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of Cheminformatics* **12**, 17, doi:10.1186/s13321-020-00423-w (2020).

17 Vu, O., Mendenhall, J., Altarawy, D. & Meiler, J. BCL::Mol2D-a robust atom environment descriptor for QSAR modeling and lead optimization. *J Comput Aided Mol Des* **33**, 477-486, doi:10.1007/s10822-019-00199-8 (2019).
18 Percepta-Batch (Advanced Chemistry Development, Inc. (ACD/Labs), Toronto, ON, Canada, 2023).
19 Marvin (ChemAxon, Ltd, 2023).
20 Molecular Operating Environment (MOE) (Chemical Computing Group, Inc., Montreal, Quebec, Canada, 2023).
21 Mansouri, K., Grulke, C. M., Judson, R. S. & Williams, A. J. OPERA models for predicting physicochemical properties and environmental fate endpoints. *J Cheminform* **10**, 10, doi:10.1186/s13321-018-0263-1 (2018).
22 de Souza Neto, L. R. *et al.* In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. *Frontiers in Chemistry* **8**, doi:10.3389/fchem.2020.00093 (2020).
23 Jiménez-Luna, J., Grisoni, F. & Schneider, G. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence* **2**, 573-584, doi:10.1038/s42256-020-00236-4 (2020).
24 Foulon, C. *et al.* Determination of ionization constants of N-imidazole derivatives, aromatase inhibitors, using capillary electrophoresis and influence of substituents on pKa shifts. *Journal of Chromatography A* **1035**, 131-136, doi:https://doi.org/10.1016/j.chroma.2004.02.053 (2004).
25 Hinman, R. L. & Lang, J. The Protonation of Indoles. Basicity Studies. The Dependence of Acidity Functions on Indicator Structure. *Journal of the American Chemical Society* **86**, 3796-3806, doi:10.1021/ja01072a040 (1964).
26 Hornberger, K. R. & Araujo, E. M. V. Physicochemical Property Determinants of Oral Absorption for PROTAC Protein Degraders. *Journal of Medicinal Chemistry* **66**, 8281-8287, doi:10.1021/acs.jmedchem.3c00740 (2023).
27 Wu, X. *et al.* Small molecular inhibitors for KRAS-mutant cancers. *Frontiers in Immunology* **14**, doi:10.3389/fimmu.2023.1223433 (2023).
28 Popow, J. *et al.* Targeting cancer with small molecule pan-KRAS degraders. *bioRxiv*, 2023.2010.2024.563163, doi:10.1101/2023.10.24.563163 (2023).
29 Drummond, M. L., Henry, A., Li, H. & Williams, C. I. Improved Accuracy for Modeling PROTAC-Mediated Ternary Complex Formation and Targeted Protein Degradation via New In Silico Methodologies. *J Chem Inf Model* **60**, 5234-5254, doi:10.1021/acs.jcim.0c00897 (2020).
30 Dixon, T. *et al.* Predicting the structural basis of targeted protein degradation by integrating molecular dynamics simulations with structural mass spectrometry. *Nature Communications* **13**, 5884, doi:10.1038/s41467-022-33575-4 (2022).
31 Zdrazil, B. *et al.* The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research* **52**, D1180-D1192, doi:10.1093/nar/gkad1004 (2023).
32 Sadowski, J. & Gasteiger, J. From atoms and bonds to three-dimensional atomic coordinates: automatic model builders. *Chemical Reviews* **93**, 2567-2581, doi:10.1021/cr00023a012 (1993).
33 Mendenhall, J. & Meiler, J. Improving quantitative structure-activity relationship models using Artificial Neural Networks trained with dropout. *J Comput Aided Mol Des* **30**, 177-189, doi:10.1007/s10822-016-9895-2 (2016).
34 Ropp, P. J., Kaminsky, J. C., Yablonski, S. & Durrant, J. D. Dimorphite-DL: an open-source program for enumerating the ionization states of drug-like small molecules. *Journal of Cheminformatics* **11**, 14, doi:10.1186/s13321-019-0336-9 (2019).