A universal model for drug-receptor interactions

Filipe **Menezes** [1,2#*], Adam **Wahida** [3#*], Grzegorz M. **Popowicz** [1,2,4,#*]

[1]*Institute of Structural Biology, Helmholtz Munich, Neuherberg, Germany;* [2]*TUM School of Natural Sciences, Department of Bioscience, Bayerisches NMR Zentrum, Technical University of Munich, Garching, Germany;* [3]*Institute of Metabolism and Cell Death, Helmholtz Munich, Neuherberg, Germany;* [4]*Khumbu.Ai GmbH;* *these authors contributed equally: F. M., A. W., G. M. P.*;* #correspondence should be addressed to F. M. (filipe.menezes@helmholtz-munich.de), A. W. (adam.wahida@helmholtz-munich.de), G. M. P (grzegorz.popowicz@helmholtz-munich.de)

The genomic landscape of disease holds, in principle, the information required for rational therapeutic design. Genes encode proteins whose functions are tightly coupled to their three-dimensional structures. Since the late 1970s, the advent of macromolecular crystallography inspired the notion that structural knowledge alone could enable a "lock-and-key" approach to drug design. However, this framework has failed to catalyze a step-change in the generation of new chemical matter. Drug discovery continues to depend on costly and largely serendipitous screening campaigns. Only by accident do we discover adaptive, dynamic receptor-drug interactions that allow for rational design. While structural biology has dramatically improved our molecular resolution, it has not solved the optimization problem, nor has it circumvented the attrition inherent in (pre)clinical development. Compounding this is the scarcity of novel chemistry beyond the reach of existing patents. At the heart of these limitations lies a fundamental bottleneck: the lack of models capable of learning and generalizing drug-target interactions to truly novel chemical matter. This work addresses that gap through a theoretical framework for predictive molecular recognition.

Introduction

Classical physical theories provide only coarse approximations of drug-target interactions, which typically unfold in two stages: first, the insertion of a ligand into the protein's binding pocket; second, the assessment of binding strength and its downstream functional consequences. The inherent, dynamic complexity of the binding event within biomacromolecular systems often renders any physics-based models inadequate for capturing biological or clinical outcomes effectively. For instance, it is well established that binding pockets adapt to the ligands they bind to, resulting in fine, but essential, variations within binding sites, a phenomenon which we coin here as ligand imprint. Furthermore, capturing the full dynamic nature of protein-ligand interactions requires simultaneously highly efficient and accurate methods, which are neither satisfied by classical force fields–fast but inaccurate–nor quantum mechanical methods–accurate but slow. While machine learning approaches excel at interpolating within chemical series, their predictive power diminishes sharply when applied to structurally diverse compounds [1,2]. Screening for new small-molecule drugs relies on large, real or virtual, compound libraries. Their member molecules were either previously synthesized or there is a predicted synthetic route to obtain them [3,4]. As a result, even the most extensive libraries are confined mainly to the historical chemical space, dominated by compounds emerging from prior synthetic campaigns or shaped by methodological bias, wherein a narrow set of reactions guides molecular design. This constraint markedly reduces the likelihood of discovering novel chemical matter with both functional relevance and intellectual property value.

Together, the true potential for structure-guided drug discovery remains untapped. Current computational approximations are often seen as a "poor man's approach" to the otherwise slow and costly Design-Make-Test-Analyze (DMTA) drug discovery cycle employed by the pharmaceutical industry [5]. The key to improving the situation lies in effectively linking the receptor's structure to the predicted ligand's binding pose and affinity, effectively increasing the DMTA turnover. Simultaneously transitioning from a biased IP-crowded to an IP-free chemical space is necessary, which largely remains unexplored [6]. Here, we propose a theory-driven machine learning (ML) model that generalizes non-bonded drug-receptor interactions by abstracting away chemical identity and learning instead spatial interaction fields, effectively modeling physical constraints that determine drug-receptor interactions from empirical structure-function observations.

A machine learning model to generalize non-bonded chemistry

Here, we report an ML model able to generalize non-bonded interactions between a drug and its targeted receptor (**Figure 1A**). The model is specifically designed to circumvent known limitations of structure-based drug discovery methods. We show that the model's predictions align well with medicinal chemistry principles and, following them, lead to improved potency. ML models were shown to be capable of deducing the behavior of physical systems using a limited set of system state parameters [7,8]. In other words, they effectively discovered their own "physics" describing

the observed phenomenon. To avoid memorization, we hypothesized that the model must not see the drug's chemical structure. This is because the landscape of drug-like molecules, estimated to encompass over 10^{60} entities, is explored by only *ca.* 20,000 published protein-ligand complex structures [9,10,12]. Instead, we trained the model on the drug's electronic atom type at a given receptor volume element (**Figure 1B**). Elements (*e.g.*, carbon), their electronic configurations (*e.g.*, C_{sp3}; **Figure 1C**), and their surrounding receptor amino acids constitute a microenvironment that is used as a training point (**Figure 1D**). The trained model infers the optimal electronic atom type at any given point in the receptor's binding site (**Figure 1E**).

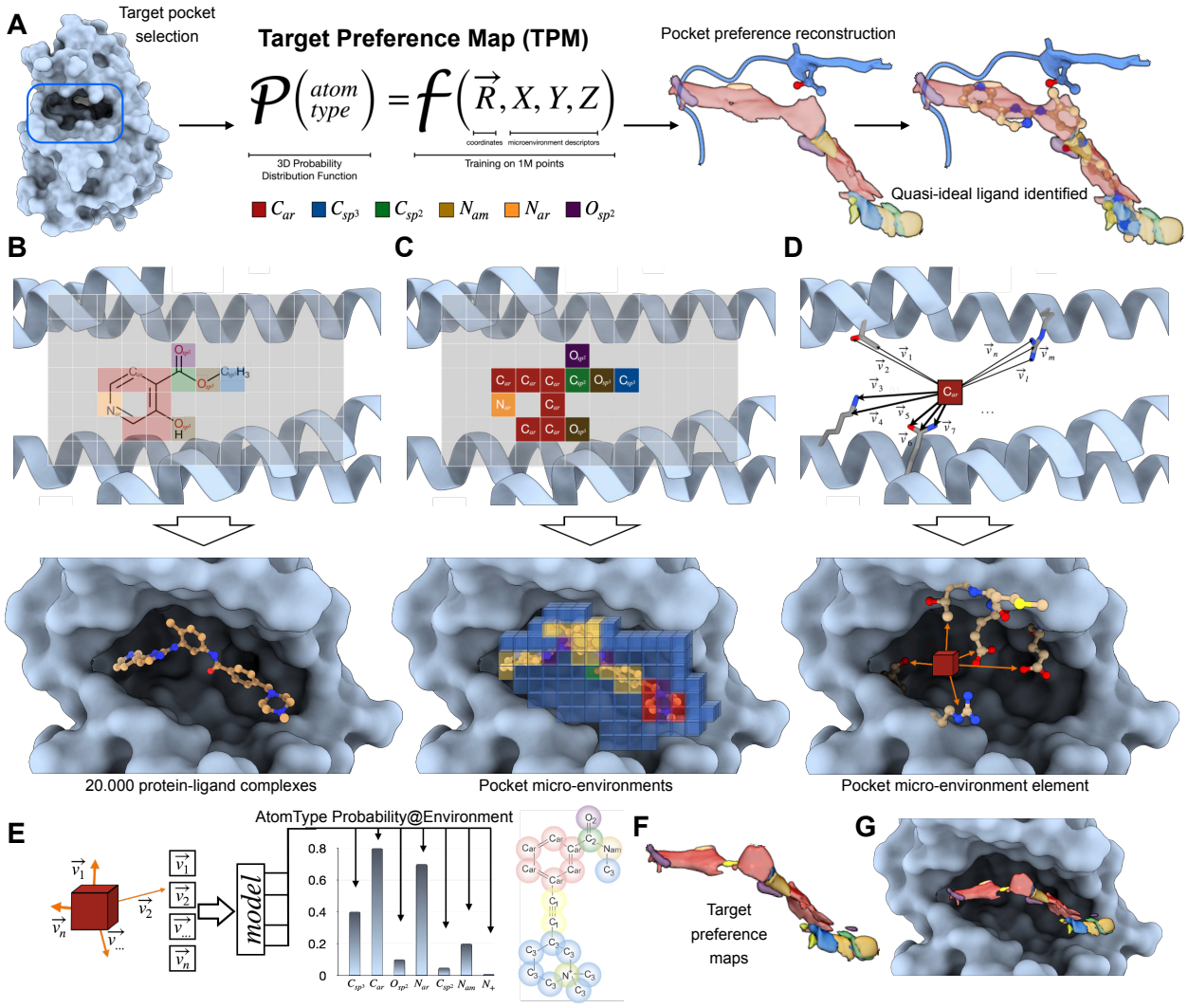


Figure 1: The Target Preference Map (TPM) Model. A) Conceptualization of the TPM model. Given the pocket of a biological target, we define atom-type-specific probability distribution functions. The latter takes as arguments Cartesian coordinates and 3D chemical microenvironments and returns the likelihood of finding specific chemical matter in a biological target's pocket. Reconstructing the ideal pockets allows us to identify quasi-ideal ligands. B-D) Two (up) and three-dimensional (down) representation of the TPM-model construction. B) The biological target's pocket is first voxelized, and then C) the atom types found in each element are identified. D) We then characterize the microenvironments by collecting all the nearby protein atoms in the form of neighbor vectors. E) The microenvironments are given as input to a transformer/perceiver model, which directly learns chemical matter distribution functions. F) When assembled in three-dimensional clouds, G) we generate target-specific localized preference functions.

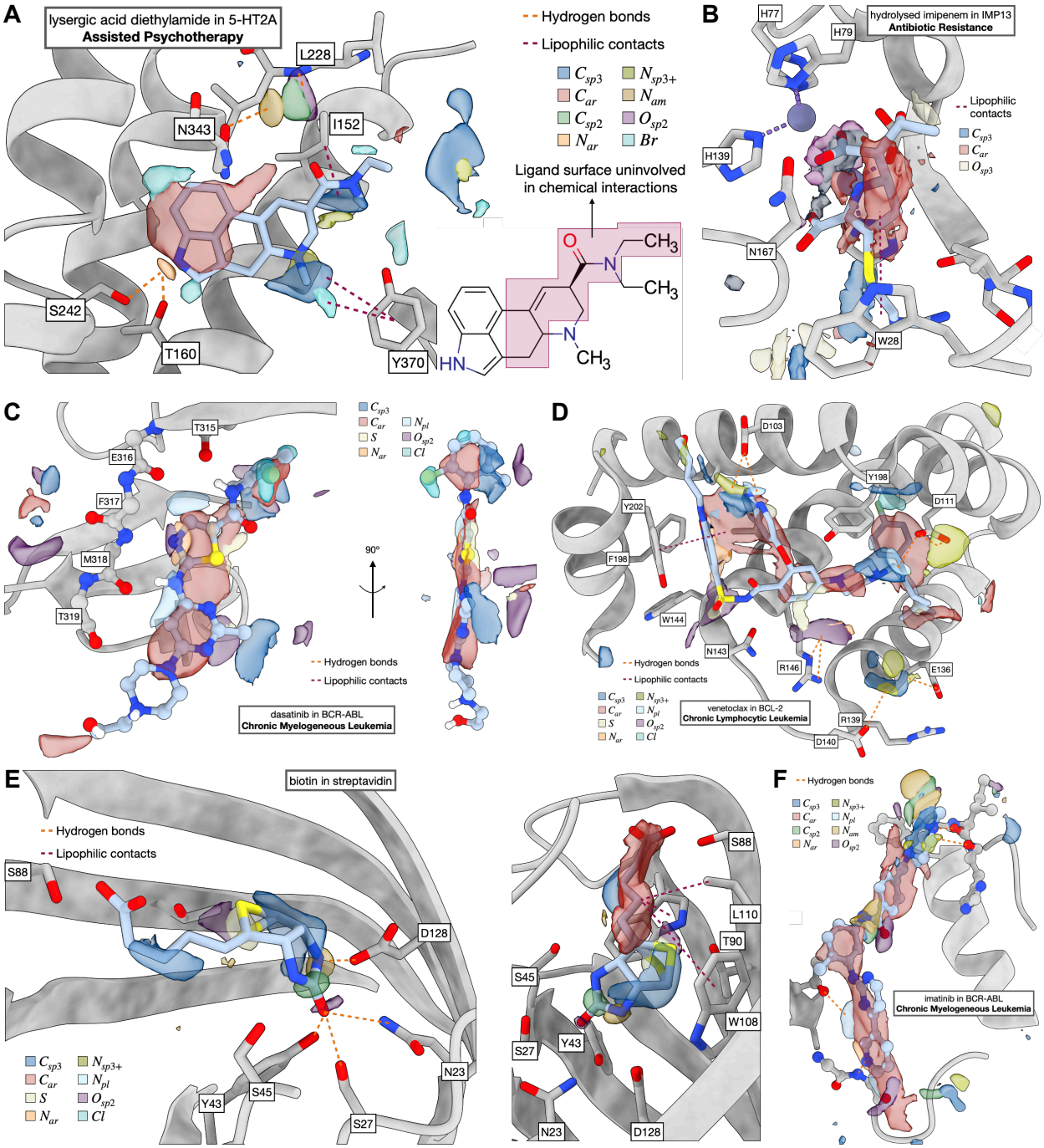


Figure 2. Gallery of TPM examples. A. Lysergic acid diethyl amide in the 5-HT2A G-Protein Coupled Receptor (GPCR), a model for psychologically assisted therapy. TPM clouds predicted for the structure of the 5-HT2A GPCR with serotonin inside (PDB-ID XYZ). Eight different clouds are represented in the structure. We observe a good overlay of the aromatic clouds with the ligand's aromatic groups. Similar observations apply to the aromatic nitrogen and the sp3 ammonium group. Lysergic acid's amide group is uninvolved in any interaction with the protein, and this is reflected in the respective TPM predictions. Lastly, one of the bromine signals matches that of the bromolysergic acid diethyl amide ligand of this GPCR. B) Hydrolyzed imipenem in IMP13 (PDB-ID XYZ). Although the model is oblivious to Zn atoms, it could understand the side chain arrangements around them and propose a carboxyl group to coordinate the Zn2+ cations as seen in known inhibitors. C) A type I kinase inhibitor, Dasatinib, in its biological target BCR-ABL (PDB-ID 2GQG), a model for chronic myelogenous leukemia (CML). A good overlay of the aromatic

clouds with the ligands' aromatic groups and identification of hydrogen bond donors and acceptors to capture the kinase's hinge region. Interestingly, the model overlays aromatic nitrogens and sp2 oxygens, despite their chemical incompatibility. This reflects the equivalence of these atom types in their hydrogen-bond accepting character. D) Venetoclax in BCL-2 (PDB-ID 6O0K), a model for chronic lymphocytic leukemia (CLL). The TPM model suggests adding ammonium groups close to some of the receptor's asparagines (D103 and D111). Aromatic clouds are meant to capture π-stacking interactions with residues Y202 and Y198. E) Biotin in Streptavidin (PDB-ID XYZ). The model predicts adding aliphatic carbons, amide groups, and sulfur atoms in good agreement with the ligand. F) TPM predictions for imatinib in BCR-ABL's pocket (PDB-ID 2HYY), another model system for CML. TPMs correctly capture the aromatic extent of the ligand, the amide group, and place aromatic and planar nitrogens to capture the kinase's hinge by hydrogen bonding.

In consequence, training becomes agnostic to ligand chemistry and historical bias is minimized. Further, abstracting chemical connectivity and exploring predictions for sub-atomic microenvironments reduces limitations stemming from a shortage of experimental data: the 20,000 protein-ligand complexes [9,10] yield about 1,100,000 training points. This enables the robust training of the model. A continuous assembly of such microenvironment predictions creates a volumetric map that characterizes a receptor's preference for drug chemistry (**Figure 1F**). We label these probability densities as Target Preference Maps (TPMs), translating the biomolecular receptor's structure into its drug preference. TPMs cover the entire investigated volume of the drug-receptor binding site (**Figure 1G**). TPMs of all atom types together predict the ideal ligand, satisfying all possible interactions the receptor could request. No real molecule can meet all, often exclusive, features. However, we observe that TPM predictions correctly reflect non-bonded chemistry rules and accurately predict critical binding features of real drugs. The drug molecules satisfying more TPM features are inherently improved binders. We show that TPM predicts non-trivial features critical for drug binding, like cofactors or bridging water molecules. In a prospective validation experiment, we demonstrate that using TPMs can overcome drug optimization problems. **Figure 2** shows TPM predictions for several receptor classes of biological targets, showcasing the generality of our approach.

In our study, we consider the following atom classes: B–boron; C_X–X takes the values *sp*, sp^2, and sp^3 for specific atomic hybridizations and *ar* for aromatic carbons; N_X–X takes similar values as carbon types, with the additional classes *am* for amide nitrogens, *pl* for non-amide planar nitrogens; O_X–X takes values similar to carbon and nitrogen hybridization states; P–phosphorous; S_X–X takes values similar to oxygen types; F, Cl, Br, I–halogen atom types, respectively, fluorine, chlorine, bromine, and iodine.

The TPM model abstracts non-bonded interaction chemistry

We trained the TPM model on 30 ligand-specific atom types frequently observed in drug molecules. Our initial benchmark was to determine whether the model abstracts the chemistry underlying non-bonded interactions from the training data. We begin by evaluating the model's accuracy (**Figure 3A**) and AUC (**Figure 3B**) relative to sample size for each atom type. Surprisingly, electronic atom types' metrics do not correlate with sample size (**Figures 3A**, **B**). Most atom types have good accuracy (>70%), with phosphorus and boron reaching excellent values (>90%), the latter nearing 95%, which may suggest overfitting due to a small sample size (~200 samples) due to nearly identical structural properties needed to bind them (e.g. hydrolytic enzymes frequently targeted by boronic acid warhead). AUC values are also overall high (>75%), except for S_{sp3} and N_{tr}. Again, boron and phosphorus score the highest, with phosphorus showing a slightly better AUC. The poor correlation between accuracy and sample size prompted further investigation. A comparison between O_{sp2} and O_{sp3}, which have similar sample sizes, shows that O_{sp2} is predicted 5% more accurately. Both atom types are involved in polar interactions, but O_{sp2} has typically lower partial charges (–0.5e vs. –0.25e), favoring stronger, more directional contacts. Likewise, C_{ar} and N_{ar} show identical accuracy despite a 10-fold difference in sample size (20,000 vs. 200,000). Here, N_{ar} engages in directional hydrogen bonding, while C_{ar} favors less directional, lipophilic contacts and π-stacking. A final example contrasts C_{ar} and C_{sp3}, which have equal sample sizes. However, the former is predicted ~5% more accurately, likely because C_{sp3} lacks π-stacking ability and only participates in non-directional lipophilic interactions. These comparisons suggest that model performance is driven more by chemical interaction types than by data quantity: atom types involved in polar and hydrogen bonding tend to be predicted more accurately. Conversely, those atom types predominantly involved in lipophilic interactions show lower performance. Thus, the sample size used was seen as adequate to train the model, and the predicted model accuracy is in fact related to the nature of chemical interactions.

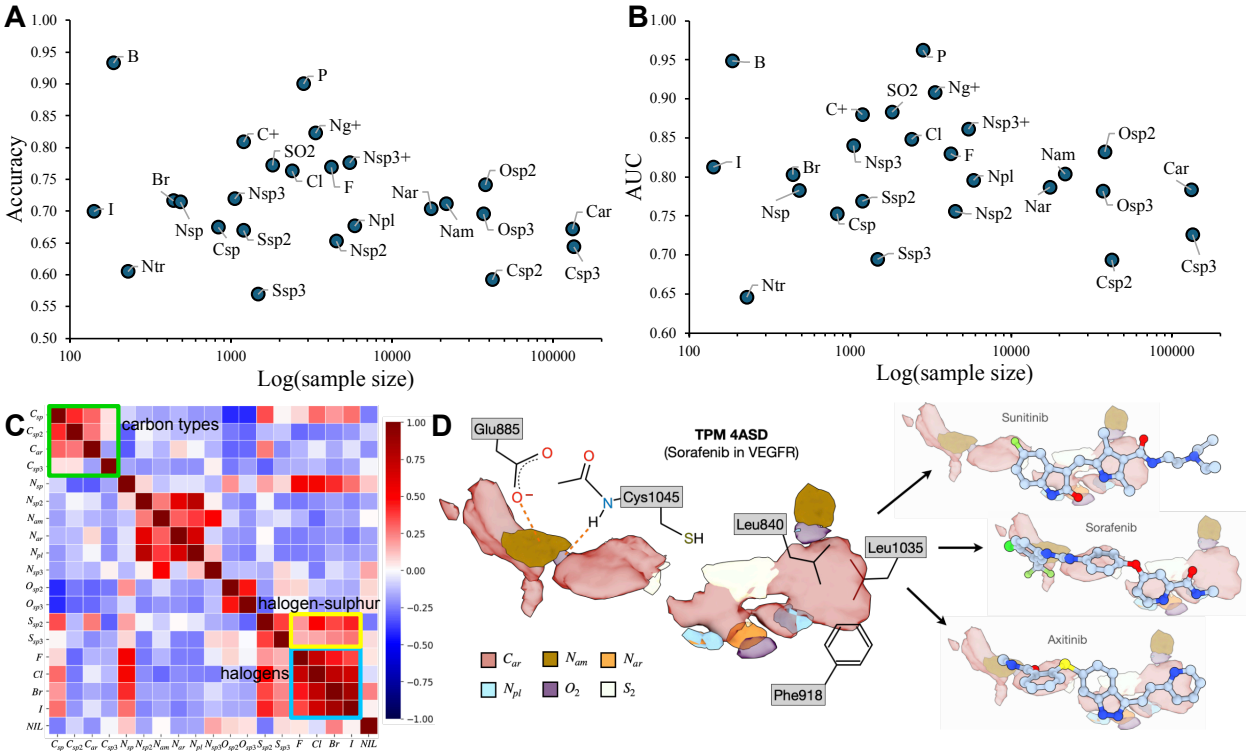


Figure 3. Statistics and Inference of the TPM model. A. TPM-model accuracy for every atom type considered against the respective sample size on a logarithmic scale. B. TPM-model AUC for every atom type considered against the respective sample size on a logarithmic scale. These two blocks show that polar atom types may achieve higher accuracy than apolar ones given similar sample sizes, because non-bonded interactions are more directional. C) Correlation matrix of atom types: given that one atom type is found at a specific point in the target's pocket, what is the likelihood of finding other, alternative atom types? D) Cross-validation of the TPM model on VEGFR with sorafenib (PDB-ID 4ASD), a model for lung cancer. Overlay of TPM predictions with the structure of other VEGFR inhibitors, sunitinib and axitinib. The TPMs show a good overlay with the chemical composition of axitinib, although the kinase's pocket contained sorafenib. Noteworthy are the predictions to extend the aromatic density and the introduction of a sulfur atom, where axitinib contains a linker.

An additional evaluation metric results from suggesting atoms of similar properties in the exact same location. We plotted the correlation between model-identified atom types for each point (**Figure 3C**). The binding-determining properties of electronic atom types correlate very well. For example, all unsaturated carbon types (C_{sp2}, C_{sp}, C_{ar}) correlate with one another, but they do not correlate with saturated carbons (C_{sp3}), devoid of π-electron systems. Expectedly, halogen atoms (F, Cl, Br, I) showed a high internal correlation. The small, electron-hard fluorine, which is the halogen atom least prone to forming dispersion (van der Waals) interactions, shows a weaker correlation with later halogens (Br, I). Conversely, the iodine correlations are poorer with earlier halogens (F, Cl), which is rationalized given iodine's higher lipophilicity. These changes in correlation are, however, gradual, such that chlorine and bromine adopt an intermediary dual behavior of potentially replacing their two neighboring halogens. Moreover, the Br-I interchangeability seems to be the strongest among halogens, reflecting the increased lipophilicity of these atoms. Halogens also correlate with sulphur atoms, again reflecting the high lipophilic character of these atom types. They also correlate well with N_{sp} as those found in nitrile moieties, which can be interpreted as if it were a fifth halogen atom; Particularly, nitrile and fluorine groups are commonly and interchangeably used in drugs to fill small lipophilic pockets. A notable cross-correlation was found for N_{sp3} and N_{am}, indicating these atoms' hydrogen-bonding characteristics. Interestingly, we also find a correlation between O_{sp2} and N_{sp}, likely resulting from their hydrogen-bond acceptor properties. However, these two atom types are not always perfectly interchangeable due to the different lipophilicity and polarity of the C≡N and C=O bonds. The oxygen predictions show weak correlation with the NIL maps, indicating a suggestion for an empty space. This is easily attributed to the implicit prediction of water molecules binding [11].

The observed anti-correlations are also quite instructive: for example, oxygen atom (O_{sp2}, O_{sp3}) predictions showed an anticorrelation with most other atom types, particularly against carbon. This is evident as O_{sp2} and O_{sp3} commonly form (dual) hydrogen bonds, which is impossible for most carbon atoms. These examples indicate that the TPM model correctly generalizes drug-receptor interactions with detail exceeding simple pharmacophore rules (H-bond donors/acceptors, aliphatic, aromatic interactions): it recognized fine differences in the electronic properties of different atomic hybridizations and correctly identified their significance to non-bonded binding. This validates the hypothesis that our model can infer the complex phenomenon of drug binding.

TPM predictions correctly describe complete drug molecules

We next wondered to what extent the model generalizes molecular interactions or whether it is decoding ligand information imprinted by receptor-induced-fit features. To investigate this, we generated TPM maps for one receptor obtained from different co-crystals with different drugs. All examples were generated after removing the structural family from the training. We took the structures of VEGFR2 with sorafenib, axitinib, and sunitinib as reference (respectively, PDB-IDs 4ASD, 4AGC, and 4AGD; **Figure 3D**). We began by analyzing the TPM-composition resulting from the 4ASD model, the VEGFR structure co-crystallized with sorafenib. These TPMs contain features characterizing the other two, more potent inhibitors, axitinib and sunitinib. Sunitinib exhibits the least overlap with the 4ASD TPM set. This results from the fact that sunitinib addresses VEGFR's pocket differently than sorafenib. In contrast, sorafenib is a type II kinase inhibitor, binding to the inactive kinase; sunitinib is a type I inhibitor, binding to the active kinase.

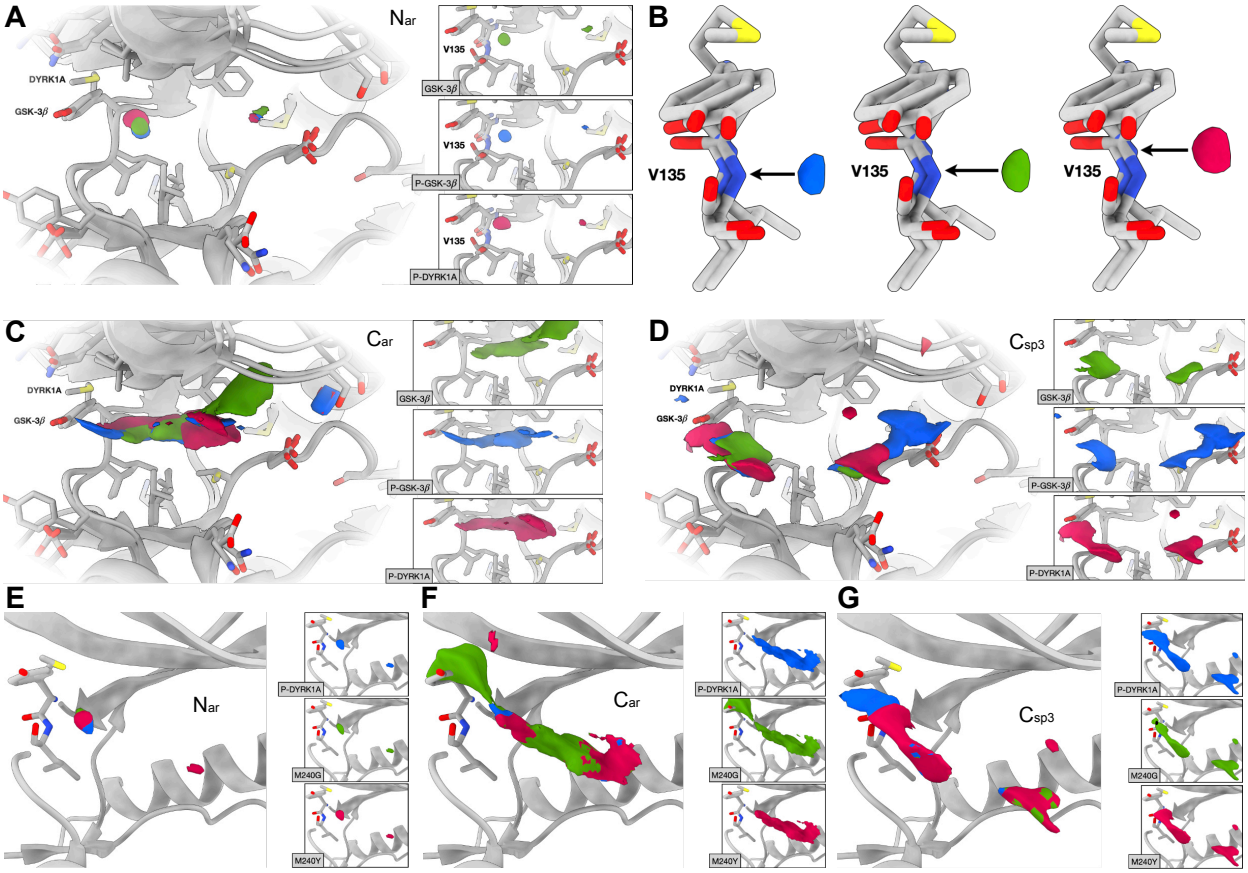


Figure 4: The TPM model understands specific non-bonded chemistry despite target similarity. In this analysis, we looked at TPM predictions for GSK3β and DYRK1A. A) TPMs for aromatic nitrogens follow the coordinates of hinge residues in GSK3β (phosphorylated and unphosphorylated, PDB-IDs XYZ and XYZ) and phosphorylated DYRK1A (PDB-ID XYZ). B) Zoom-in of the respective predictions with the kinase structures overlaid. The arrows are positioned at the geometric center of each signal. C) Aromatic carbon predictions and D) sp^3 carbon predictions for the same systems. In both cases, cloud details are system-specific, although a general shape is somewhat conserved. E-G) Modifications in TPM-cloud predictions for mutations transforming gradually DYRK1A's hinge into that of

GSK3β. E) Although aromatic nitrogen signals remain conserved, changes are recorded for F) aromatic and G) sp^3 carbons.

Despite the different kinase conformations, some structural elements are conserved, and the TPM model suggests how to explore the receptor-drug interactions best. This is, for instance, the alignment of sunitinib's N_{pl}, neutral, non-amide sp^2 nitrogen atoms with three neighbors, signal, and the pyrrole ring, along with the oxoindole part of the ligand. The oxoindole's oxygen also aligns well with the O_{sp2} signal captured from the crystal structure. Axitinib achieves an even better alignment with the TPMs. The amide signal (N_{am}) overlaps with the ligand's amide group, and the ligand provides better coverage of C_{ar} predictions. Importantly, the amide signal is also present in the TPMs calculated from sunitinib's co-crystal structure, despite the latter ligand not accessing that area of the pocket. This signal, which cannot stem from ligand imprinting, suggests that the amide functionality is vital for binding. We also observe that axitinib's indazole group aligns with the TPMs calculated from sorafenib's co-crystal structure for N_{pl} and N_{ar}, even though sorafenib only positions the N_{ar} functionality in that region. Notably, no N_{ar} signal is detected near the terminal pyridyl group of axitinib. **Figure S1A** presents a view of axitinib in the VEGFR2 pocket, showing no direct interaction between this group and the protein. This also confirms that the drug features are not memorized but inferred. Nevertheless, this ring contributes to extending the aromatic signal requested from the 4ASD-TPMs. Lastly, we emphasize that axitinib places a sulfur atom close to a sulfur (S_{sp3}) signal recorded from the Sorafenib complex structure that lacks such a sulfur linker. **Figure S1** further complements this analysis with data on the other two crystals.

Our analysis above shows how the TPM model distils critical drug binding features. This is done independently of the drug molecule in the experimental structure used for calculations, suggesting that TPM predictions are not significantly biased by memorization and are derived from drug binding generalizations. This indicates that TPMs can guide medicinal chemistry optimization towards new, more potent drugs and unrestricted chemical variability. The TPM predictions can cover the entire receptor binding site's volume, which is larger than the experimental ligand's volume. This can be directly used to guide chemical optimization. Obviously, some TPM features disagree with the baseline structure. This is evident as the idealized ligand that fulfils all TPM features is chemically impossible. This is seen in **Figure S2**, where signals for O_2 and N_{ar} overlay to exploit the same hydrogen bond interaction with a kinase's hinge. **Figure S2** further details an example of type I and type II kinase binders and the respective TPM signals. It is seen that there is good agreement between the TPMs generated from different type I inhibitors. We have performed additional TPM predictions on binding sites for standard classes of drug targets. The predicted maps show astonishing similarities with highly potent drug molecules (**Figures S3**, **S4**, and **S5**). The predictions were calculated using a model trained without the query structures to remove the possibility of data leakage and memorization effects. In conclusion, the TPM model captures the shape and molecular groups of experimentally determined structures.

TPMs systematize drug-protein interactions

Next, we tested TPM sensitivity towards fine structural differences between two, structurally similar disease-associated kinases: GSK3β and DYRK1A [13] (accession codes 7Z5N, 7Z1F, and 7Z1G; **Figure 4**). To minimize the possible ligand-imprinted fingerprint bias, we selected only structures with the same inhibitor in the pocket. The TPM signals show resemblance for all three cases. For instance, the TPM densities for aromatic nitrogen atoms (N_{ar}; **Figures 4A, B**) track the hinge residue they are hydrogen-bonded to. However, maps for aromatic and sp^3 carbons (C_{ar}, **Figure 4C**; C_{sp3}, **Figure 4D**), while maintaining the position, have vastly different shapes and extents. To test whether these subtle differences result from differing atomic coordinates of pocket residues, we mutated DYRK1A's Met240 to Gly (mutation 1) to open the binding site and to Tyr (mutation 2), to mimic GSK3β (**Figures 4E-G**). Expectedly, N_{ar} signals remained unchanged (**Figure 4E**), as these modifications cannot affect the hydrogen-bonding properties of DYRK1A's hinge region. Yet, Gly240 allows the TPM-C_{ar} cloud to expand (**Figure 4G**). Introducing a Tyr side chain, on the other hand, constricts the extent of the aromatic cloud (**Figure 4G**). This example proves that the TPMs are highly sensitive to local binding site composition and do not memorize the overall receptor-drug structure.

TPM generalizes the complex transition metal chemistry

We proved that the TPM model is sensitive to the receptor binding site's chemical nature and spatial arrangements. Due to sparse training data, we have removed co-factors from training, including metals. However, cofactors leave a prominent structural fingerprint on the receptor. This is recognized by the model and generalized. To investigate this, we examined the impact of transition metals, which the TPM model is intrinsically agnostic to, on the TPM-density predictions. We selected the complex case of the IMP-13 metallo-β-lactamase, an important antibiotic-resistance factor and drug target lacking approved medications. The IMP-13 coordinates two adjacent zinc ions at its active site.

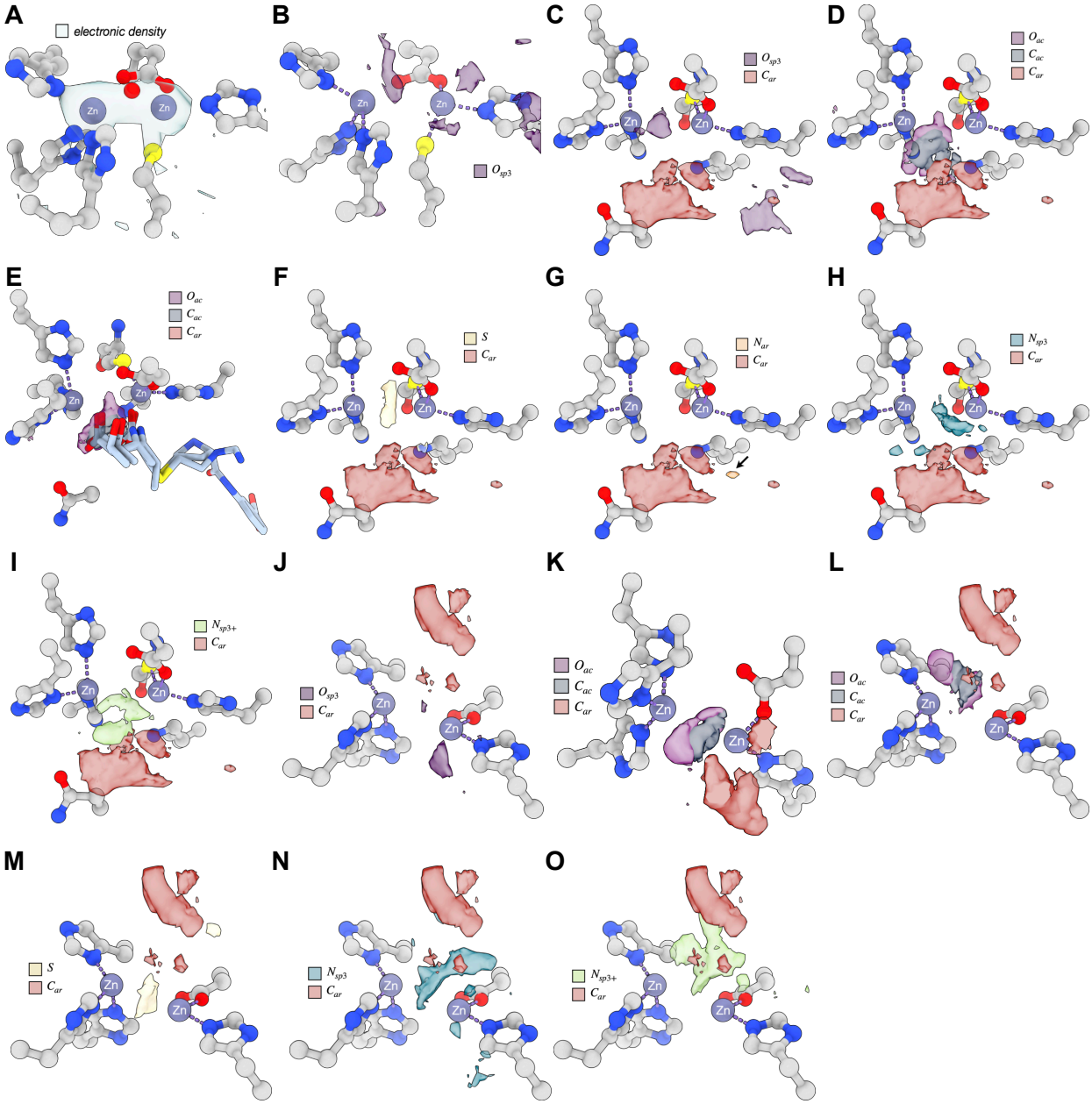


Figure 5. TPM predictions of the IMP13 metalloenzyme. A. The experimental electronic density of apo IMP13 (PDB-ID 6R78) showing the presence of an oxygen atom between the zinc cations (hydroxy group). B-I) TPM predictions for IMP13 with hydrolysed meropenem (PDB-ID 6R73). B) sp3 oxygen signals overlay with good accuracy with the bridging oxygen atom between the zinc atoms, even though the structure used to generate TPMs was not of the apo protein. C) Alternative view of the predictions in the previous block, showing the relative position of oxygen atoms between the zinc ions. Inclusion of aromatic signals. D) Prediction of carboxyl groups and E) overlay of those signals with the structures of several inhibitors containing the carboxyl group. F) Predicted sulfur clouds. G) Predicted aromatic nitrogen clouds. Although aromatic nitrogens and sulfur atoms share some properties, e.g., binding to metal ions, the predictions are atom-type specific. On the other hand, H) amine nitrogens are sought between the zinc cations. I) The TPM prediction for ammonium groups is more scattered. Although these atom types are incompatible with zinc cations, they both have similar chemical characters. J-O) Similar predictions for IMP13 with hydrolyzed doripenem (PDB-ID 6S0H).

These zinc atoms activate water molecules to form hydroxyl anions, responsible for hydrolyzing beta-lactam antibiotics. This is seen in the crystal structure of apo-IMP-13, PDB-ID 6R78, **Figure 5A**. This mode of catalysis cannot be classically rationalized. TPM predictions using the crystal structure with PDB-ID 6R73, which is co-crystallized with hydrolyzed meropenem, show a prominent O_{sp3} signal precisely where the hydroxyl anion bridges the two metal centers (**Figures 5B, C**). We stress that, in 6R73, there is no bridging hydroxyl group between zincs, as the hydrolyzed antibiotic sterically and chemically hinders hydroxyl coordination. A prominent TPM density for carboxylates appears at the binding site, which almost perfectly overlaps with the carboxylate groups in other ligand-bound structures (6RZR, 6RZS, and 6S0H; **Figures 5D** and **5E**). We stress that, since the cocrystal 6R73 does not contain Zn-bound carboxylates, these signals cannot result from ligand imprint. It is also worth noting that, although TPM training discards any connectivity information, the model recovers the structural features of the carboxylate group. Notably, the TPM model also suggests the introduction of sulfur-containing functional groups in proximity to the zinc ions (**Figure 5F**), which aligns with the coordination chemistry of zinc. Interestingly, no N_{ar} density is predicted around the metal centers (**Figure 5G**). We speculate that this absence reflects the spatial separation of the N_{ar}/C_{ar} density from the zinc sites, such that introducing aromatic nitrogens provides no additional stabilizing interaction. Another notable signal is the strong TPM density for ammonium groups (N_{sp3}, **Figures 5H** and N_{sp3+}, **Figure 5I**), extending close to the zinc ions. Although ammonium ions cannot coordinate transition metals, they share their Lewis acidic character with metal cations, which may underlie this unexpected similarity in TPM signals. This is, in our experience with the TPM model, one of the rare examples where N_{sp3}, N_{sp3+}, and sulphur signals overlay with such precision. To further corroborate these findings, TPM densities were calculated for an additional structural model, yielding consistent results and reinforcing our conclusions (**Figures 5J-O**).

TPM leapfrogs constraints in conventional medicinal chemistry optimization

The TPM predictions, when compared to known drug placement, guide towards better binders. However, every retrospective analysis can be flawed by unconscious human bias. Therefore, we tested the hypothesis progressively by synthesizing molecules binding to a difficult-to-target protein-protein interaction interface. We chose the protists' PEX14-PEX5 interface for medicinal chemistry development in our lab, as it proved extremely difficult (**Figure 6a)** [14,15]. Classical optimization reached an apparent plateau: conventional structure-based drug design strategies and traditional medicinal chemistry approaches failed to yield further improvements in its pharmacological properties (EC_{50} (*T. cruzi*) = 0.851 μM). To overcome this barrier, we generated TPMs from the PDB-ID 5L87 and overlaid them with a docking pose of the lead compound **1** (**Figure 6B**). In result, we identified two non-obvious and thus previously overlooked opportunities for structural modification (**Figures 6B, C**): 1) On the left side of the binding pocket a C_{sp3} patch was suggested, which can be satisfied by introducing an alkyl group (*e.g.*, methyl, ethyl, propyl, or cyclopropyl) onto the PEX14-inhibitor western aromatic moiety (Ar^W). 2) Examining the C_{ar} density indicated that the eastern moiety (Ar^E) attachment to the central scaffold was suboptimal, as substantial parts of this moiety did not overlap with the TPMs properly. To enhance fit, the attachment point of Ar^E to the central scaffold can be shifted by a single atom, resulting in a regioisomer. Beyond these proposals, the TPMs predicted a patch of N_{sp3+} density close to the negatively charged residues Glu34 and Asp38, suggesting the potential for hydrogen bonding or salt bridge interactions. Notably, this observation was already experimentally validated in our previous structure-activity relationship (SAR) studies, where introducing an ethylammonium group to the pyrazolo[4,3-*c*]pyridine core of PEX14 inhibitors significantly enhanced activity [15]. Note that the ligand co-crystallized with PEX14 in the PDB structure 5L87 does not contain such a chemical functionality, evidencing model generalization. Following the TPM indications, we synthesized four new inhibitors, compounds **2-5** (**Figure 6D**), and assessed their efficacy against *T. cruzi in vitro*. Strikingly, most TPM-suggested modifications contributed to a substantial boost in potency. Compound **3**, in which the lead compound **1**'s chloro substituent was replaced with a cyclopropyl group, reduced parasitic growth nearly fivefold more effectively while also achieving a significantly improved selectivity, selectivity index (SI) of 42.1 vs. 13.4; SI = EC_{50} (*T. cruzi*)/CC_{50} (L6 skeletal muscle cells). Interestingly, this effect was not observed for compound **2**, which contained a methyl group in place of the chloro substituent. Still, compound **2** displayed nearly identical cellular activity to the lead compound **1** (EC_{50} (*T. cruzi*) = 0.834 μM). This outcome is nonetheless consistent with the TPM predictions, as the methyl group does not extend sufficiently to overlap with the C_{sp3} density maps. Modifying the core scaffold of a lead compound, a strategy known as scaffold-hopping, is often synthetically demanding, as it requires substantial alterations of synthetic routes. Therefore, constructive guidance is essential before the effort is made. Following the TPMs, we generated regioisomers that exhibited a consistent increase in activity. The regioisomer **4** demonstrated a twofold enhancement in pharmacological potency relative to inhibitor **1**, accompanied by superior alignment with the generated TPMs. This conformational adjustment also optimizes the ligand's ammonium group placement, which should better overlap with the predicted N_{sp3+} density, situated between the two negatively charged residues.

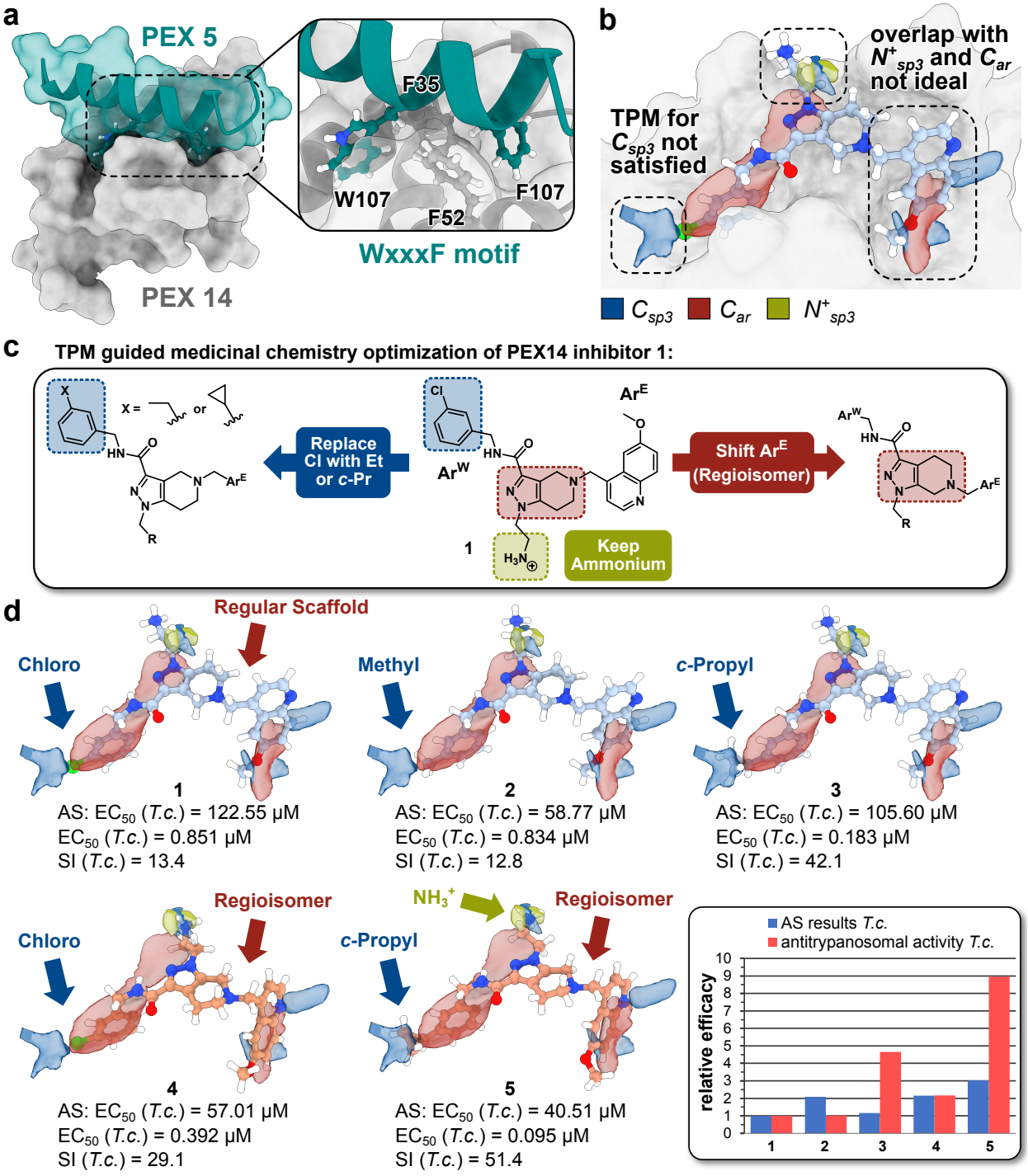


Figure 6: Prospective validation of the TPM model on a challenging protein-protein interaction (PPI) target. A) The PPI interface of PEX14 and PEX5, and the main residues involved in the interaction. B) TPM predictions based on the 5L87 crystal structure, containing PEX14 bound to one inhibitor. The maps reveal three possible structural modifications: i) introduction of alkyl groups on the leftmost aromatic substituent (also termed western aromatic moiety); shift of the rightmost aromatic group (also termed eastern aromatic moiety) to match the aromatic density prediction; and introducing an ethyl ammonium group in the center of the ligand to capture the residue XYZ. Note that the ligand in the reference crystal structure does not contain the ammonium group. C) Medicinal chemistry

strategy used for preparing the new inhibitors. D) In vitro efficacy of the new inhibitors, along with their target selectivity.

Intuitively, an ammonium group between Glu34 and Asp38 facilitates hydrogen bonds and salt bridges. Regioisomerisation is also expected to improve compound **4**'s quinoline moiety (Ar^E) alignment with the C_{ar} TPM compared to derivatives containing the regular pyrazolo[4,3-*c*]pyridine scaffold (compounds **1-3**). Finally, the target molecule **5**, which incorporates all previously established modifications and exhibits the best overlap with the generated TPMs, emerged as the most potent inhibitor across all biological assays, enhancing trypanocidal activity by nearly tenfold, compared to the previous lead compound **1**. These findings highlight the exceptional predictive accuracy of the TPM model, which is an essential feature for drug design, often lacking in current state-of-the-art *in silico* tools. Moreover, this result exemplifies why such a modification would likely be overlooked in conventional lead optimization campaigns. This result highlights the practical value of the TPM approach in drug design, mainly as it directs researchers toward unconventional and subtle medicinal chemistry modifications that will otherwise remain unexplored. We also performed a quantum mechanical evaluation of the new series of inhibitors to identify a physically motivated rationale behind the improved efficacy. The results, detailed in Figure S5, show an excellent correlation between TPM-suggestions, experimental affinity, and the quantum mechanical evaluations, providing yet additional validation of the model.

Discussion

We have demonstrated that an ML model can correctly learn and infer from principles of non-bonded chemistry. The TPM mapping provides a probabilistic model of an ideal drug for each receptor structure. This removes drug-receptor interactions from latent space towards an explainable, physical model. The model's sensitivity to minute structural detail allows calculation of drug preference changes at minimal receptor alterations like point mutations. On the general level, precision medicine aims to deliver tailored therapies by decoding the molecular underpinnings of disease. However, advances in computational biology have further revealed the predictive power of genomic data, especially when coupled with machine learning, enabling a shift from serendipitous discovery toward systematic design. In principle, the genomic blueprint of a disease harbors the information needed to guide rational drug development. Yet, current strategies, such as inferring structure from sequence, fall short of capturing the atomistic detail required for therapeutic design. A key bottleneck remains in constructing models that can learn the rules of drug-target recognition. To this end, we present an AI platform for small-molecule discovery, underpinned by a machine learning model trained on over 1 million ligand-protein interactions. This framework identifies binding determinants and optimizes drug-like properties with atomic resolution. Our integrated approach recapitulates the chemical features of existing therapeutics, enhances suboptimal leads, and accelerates de novo design against challenging targets, all with minimal design latency.

The complex nature of biological structural data and the limited diversity of training sets hinder the training of robust and reliable ML models. Complex dynamic, or quantum mechanical problems like drug-receptor binding, are delegated to the latent space. Compared with the nearly limitless chemical diversity, the experimental noise and the limits of training data render these models susceptible to overfitting and poor generalizability. Even the best-performing models suffer from heavy memorization and dramatically lose accuracy when dealing with molecules deviating from their training sets [1,2]. We shifted the focus from entire molecules to local atomic microenvironments to mitigate this. Thus, we divided the available ligand-biomolecule data into local structural features that are chemically diverse and numerous to allow the model to generalize non-bonded drug-receptor interactions. This renders the TPM model agnostic to chemical connectivity and offers a more realistic description of drug-target interactions while minimizing the bias from the historical chemical space. The model trained in this way excelled in predicting critical drug binding features. Retrospective and prospective validation analyses proved to be remarkably accurate. The TPMs are easy to read using traditional structural biology software (MRC format), which allows non-experts to interpret the maps easily. We anticipate it will be easy to use TPMs as input/conditioning for docking and molecule generation models. Docking to the TPM (receptor preferences) might be particularly efficient as each point of TPM carries chemical knowledge of all structures analyzed instead of relying on a small set of force-field functions.

The proposed approach comes at the cost of completely neglecting affinity data. This was intentional, as affinity data is noisy, unreliable, and indivisible from the ligand's structure. Instead, we assumed that observing an atom in a structural experiment motivates its inclusion in the training set. Naturally, atoms that barely contribute to binding but are present in the molecules contain unwanted noise, but their elimination would require atom-specific energy contribution calculations [9]. TPMs can be directly interpreted by humans or used in structure-aware software for

docking or *de novo* ligand generation. The TPM densities proposed by the model can help medicinal chemists to direct synthetic plans towards more rational molecules. In ML, they can serve as conditioning for diffusion-based molecule generators. TPMs can also work in connection with molecular dynamics and quantum chemistry calculations to finally realize the structure-based drug discovery paradigm. Our model now functions as a foundation for rethinking computational drug discovery and accelerating the development of urgently needed therapies. Tailoring therapeutics to a patient's genotype [16] necessitates a flexible tool to connect genetic information with drug chemistry interactions, which remains to be achieved. The use of atom-focused microenvironments by an ML or quantum computational system might lead to it being faster than massive training campaigns on heavily biased data.

Limitations

While TPM offers a novel framework for generalizing non-bonded interactions, several limitations merit consideration. First, the model operates purely at the level of atomic microenvironments and is agnostic to chemical connectivity. This abstraction, while beneficial for generalization, sacrifices explicit information about intramolecular interactions such as strain, torsion, or conformational energetics that are critical during compound optimization. Secondly, the current TPM framework is trained exclusively on static, experimentally resolved structures, which do not capture receptor dynamics, allosteric effects, or induced fit. As a result, temporal aspects of binding and entropic contributions remain outside the model's predictive scope. This is mitigated when receptor is structurally pre-conditioned towards binding by a ligand already present in the structure (like ATP) but might pose a challenge when working with apo-structures captured at their solvent-adapted conformation. Thirdly, affinity information is intentionally excluded from training to avoid confounding structural inference with noisy or context-dependent potency data. While this enhances the model's robustness for identifying preferred interactions, it limits its direct applicability to predicting absolute binding free energies or ranking close analogs. Fourth, the current training data, though diverse, still reflect historical chemical and structural biases inherent to PDB-deposited complexes. This may constrain the generalizability of TPMs to underrepresented protein families or unconventional chemotypes. Finally, the model does not yet integrate physical forces explicitly (e.g., electrostatics, dispersion, solvation), which could enhance interpretability or performance in edge cases. Addressing these limitations, by incorporating molecular dynamics-trained TPM, energetic modeling, or multimodal learning, will be key to evolving TPM into a unified framework that bridges statistical inference with physical principles.

STAR Methods

Model construction. To predict the spatial localization of ligand atoms within protein binding sites, we developed a Transformer-based neural network that estimates, for any point in a pocket, the confidence with which a ligand heavy atom of a specified type occupies that location. The model architecture consists of three stacked self-attention blocks, followed by a one-vs-all classification head that outputs atom-type-specific occupancy confidence scores. Each query point is characterized by the 25 nearest heavy atoms from the protein structure. These atoms are one-hot encoded into one of 21 protein atom classes, defined based on a combination of PDB atom names and residue identities. Atom type classes were selected to reflect common environments in drug-like molecules, with rare or minimally represented types (*e.g.*, silicon) omitted. Ligand atom types were assigned using Open Babel conventions [17], ensuring consistency with the input featurization of protein atoms. To ensure rotational invariance, we define a local coordinate system at each query point [11]. The x-axis points toward the nearest protein atom; the y-axis is constructed by Gram–Schmidt orthonormalization of the vector to the nearest non-carbon atom, constrained to be at least 30° from the x-axis; and the z-axis completes the right-handed frame. The positions of the 25 surrounding protein atoms are then expressed in this local frame and concatenated with their atom-type embeddings. The combined spatial and chemical representations are processed by three self-attention blocks (four heads, hidden dimension 64, dropout 0.2). A global mean pooling layer summarizes the contextualized representation of the local environment, which is then passed to a feedforward classifier to yield confidence scores for each ligand atom type.

Model training. Training and evaluation data were derived from the PDBBind2020 dataset [9,10]. To prevent information leakage between training and test sets—particularly since the spatial arrangement of 25 receptor atoms can be highly informative—we clustered all receptor sequences at 70% identity using CD-HIT [18]. From these clusters, 15% were randomly selected and held out to form the test set. This ensures that no closely related proteins appear across train and test splits. For each complex, ligand heavy atoms located within 5 Å of any protein atom were identified, resulting in approximately 921,000 training and 240,000 test samples. These atoms serve as the positive targets. To train the model to distinguish true ligand positions from plausible but incorrect locations, we generate negative samples by perturbing each positive atom's coordinates with isotropic Gaussian noise. If the resulting decoy location was at least 1.8 Å from all ligand atoms, it was accepted as a negative (NIL) sample. This distance threshold avoids generating trivial negatives with obvious steric clashes, while still challenging the model to learn meaningful spatial constraints. We employed a simple train-test split without a separate validation set, as model selection did not involve early stopping or hyperparameter tuning based on performance metrics. Nonetheless, the current framework supports standard extensions such as k-fold cross-validation within the training set if required.

Training protocol. We adopted a two-stage training strategy comprising initial pretraining on ligand occupancy and subsequent per-atom-type fine-tuning. In the first stage, the model was trained to distinguish ligand-occupied sites from unoccupied ones (NIL class). A binary cross-entropy loss was used to optimize the model over all available positive and NIL samples. This pretraining established general spatial and chemical features that discriminate the ligand's presence. In the second stage, we fine-tuned the model separately for each ligand atom type with at least 100 positive examples in the training data. Each atom-type-specific classifier was initialized from the NIL model's weights and trained in a one-vs-all setup: We employed dynamic resampling to ensure effective use of all training data while avoiding class imbalance. Thus, at the start of each epoch, a random subset of negative samples was drawn to match the number of positives. Using this procedure, we trained 38 distinct atom-type classifiers. For evaluation and visualization, we focused on the subset of types with sufficient test-set representation, including: N_{sp3}, O_{sp3}, C_{sp3}, Nil, B, I, S_{sp3}, O_{sp2}, C_{ar}, S_{sp2}, N_{pl}, S_{o2}, F, C_{sp}, P, N_{ar}, Cl, N_{am}, Br, N_{sp2+}, C_{sp2}, N_{sp2}, N_{sp}. All models were trained using the Adam optimizer with an initial learning rate of 8×10^{-4}. A cosine-annealing learning rate schedule was applied: Five epochs for the NIL model, and between twenty and forty epochs for the atom-type classifiers, depending on the number of positive samples available per class.

Model Evaluation. Model performance was primarily assessed using the area under the receiver operating characteristic curve (ROC-AUC). This threshold-independent metric reflects the probability of correctly ranking a randomly chosen positive above a randomly chosen negative. We selected ROC-AUC as the principal metric because it is robust to class imbalance and does not rely on a predefined decision threshold. To ensure that the evaluation focused on meaningful atom-type discrimination, we excluded NIL (unoccupied) samples from this analysis. Each atom-type-specific classifier was therefore evaluated solely on its ability to distinguish its target atom type from all other ligand atom types in the held-out test set. For each model, predictions were computed across the entire test set, and ROC-AUC scores were calculated using only the relevant positive and negative ligand atom labels. To ensure

statistical reliability, only atom types with at least thirty positive samples in the test set were included in the reported evaluation. While other classification metrics such as accuracy, recall, and precision were computed during development, they were excluded from the final analysis due to their dependence on arbitrary thresholds and the limited interpretability for classes with few test samples. Bootstrapping would be required to report these metrics, particularly for low-frequency atom types meaningfully, and may be considered in future work.

Map Generation. Following model training, atom-type classifiers can be applied to any point within a protein pocket to estimate the confidence that a ligand atom of a specific type occupies that location. By systematically evaluating these models across a three-dimensional grid, we generate high-resolution preference maps that visualize spatial binding propensities for individual atom types. To construct these maps, the binding site is discretized into a regular grid of cubic voxels, typically spaced at 0.25 Å intervals. This resolution provides a balance between spatial accuracy and computational efficiency. At each grid point, all trained atom-type classifiers are applied to produce a corresponding confidence score. The resulting maps encode the learned structural preferences—derived from experimental protein-ligand complexes—for each atom type across the binding pocket. The maps serve as interpretable, model-derived representations of chemical complementarity, highlighting regions of the pocket that are geometrically and chemically suited to specific ligand functionalities.

Synthetic validation. PEX14's pocket lipophilicity stems from two phenylalanine amino acids, Phe35 and Phe52, which are crucial for the interaction with PEX5 via one of the several diaromatic WxxxF peptide motifs present in the intrinsically disordered N-terminal region of this import receptor protein. Inspired by this binding mode a series of small molecule inhibitors was designed and synthesized using the following main scaffold: two aromatic moieties (Ar^W and Ar^E) are linked via a pyrazolo[4,3-*c*]pyridine heterocycle (**Fig. 4c**). Through multiple iterative cycles of hit-to-lead optimization, a diverse array of aromatic moieties for Ar^W and Ar^E, including substituted naphthyl, benzyl, phenyl, pyridazine, pyridine, quinoline, and indole derivatives, was systematically investigated. This comprehensive exploration led to the discovery of over 500 PEX14 inhibitors, several of which demonstrated significant enhancements over the initial screening hit. However, with compound **1** (**Fig. 4c**), the optimization reached an apparent plateau, as conventional structure-based drug design strategies and traditional medicinal chemistry approaches failed to yield further improvements in its pharmacological properties (EC_{50} (*T. cruzi*) = 0.851 μM). Consequently, retaining this modification in the next generation of PEX inhibitors was imperative. However, the TPM model indicated that further optimization was possible, as the ammonium group of compound **1** did not achieve optimal alignment with the predicted Nsp^{3+} density maps. We hypothesized that this alignment could improve for the planned regioisomer derivatives.

Quantum Mechanical calculations. To further validate the TPM-suggested ligand modifications we ran quantum mechanical evaluations using our in-house ULYSSES software [19]. Validation and analysis was performed using quantum mechanical analysis protocols as described in previous work [20,21].

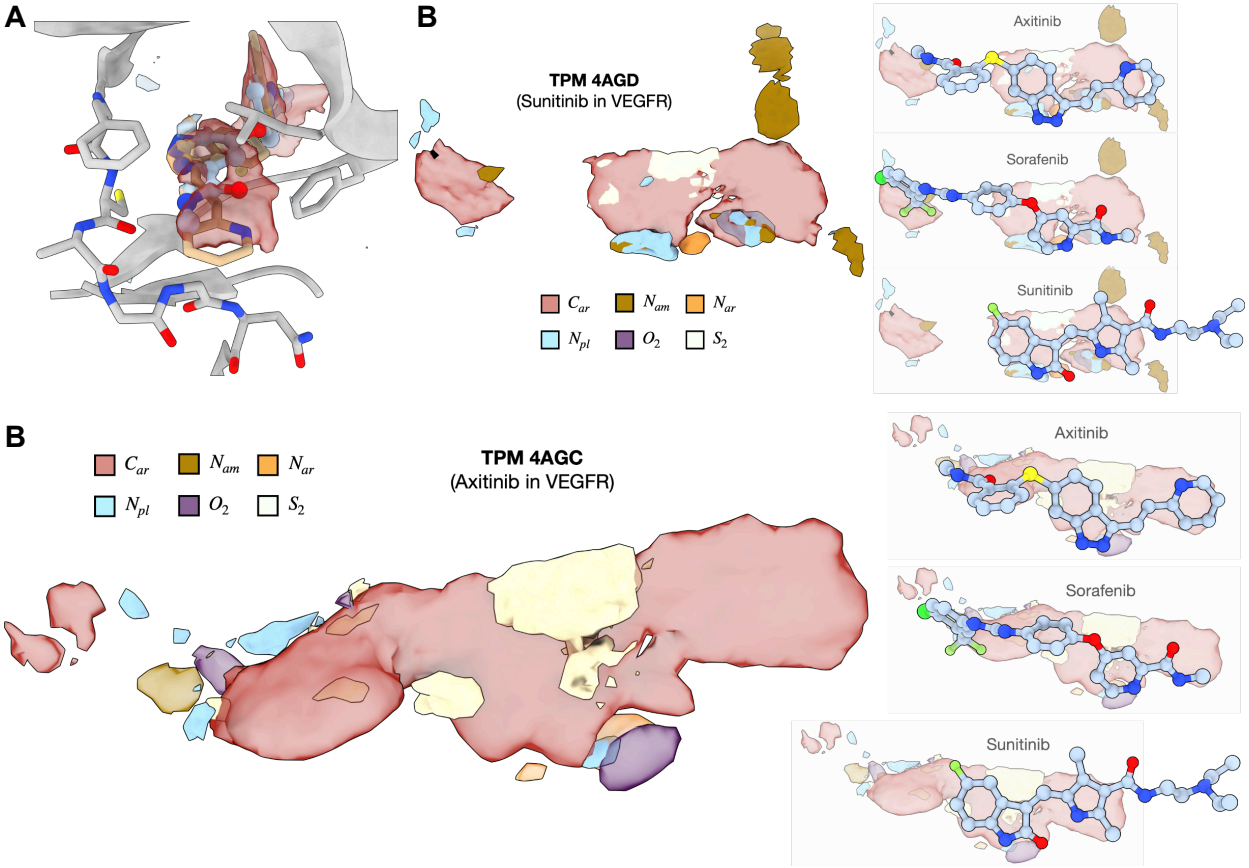


Figure S1. Additional validation of the TPM model for VEGFR crystal structures with sorafenib, sunitinib, and axitinib. A) Overlay of sorafenib and axitinib structures and the TPM predictions for sorafenib in VEGFR. Axitinib's pyridyl ring's nitrogen is not involved in any interaction with the target, and the TPM model reflects this. B) TPM predictions for the crystal structure of sunitinib in VEGFR (PDB-ID 4AGD) and overlay with the binding pose of other ligands considered in this analysis. C) TPM predictions for axitinib in VEGFR (PDB-ID 4AGC) and overlay with the binding poses of other ligands considered.

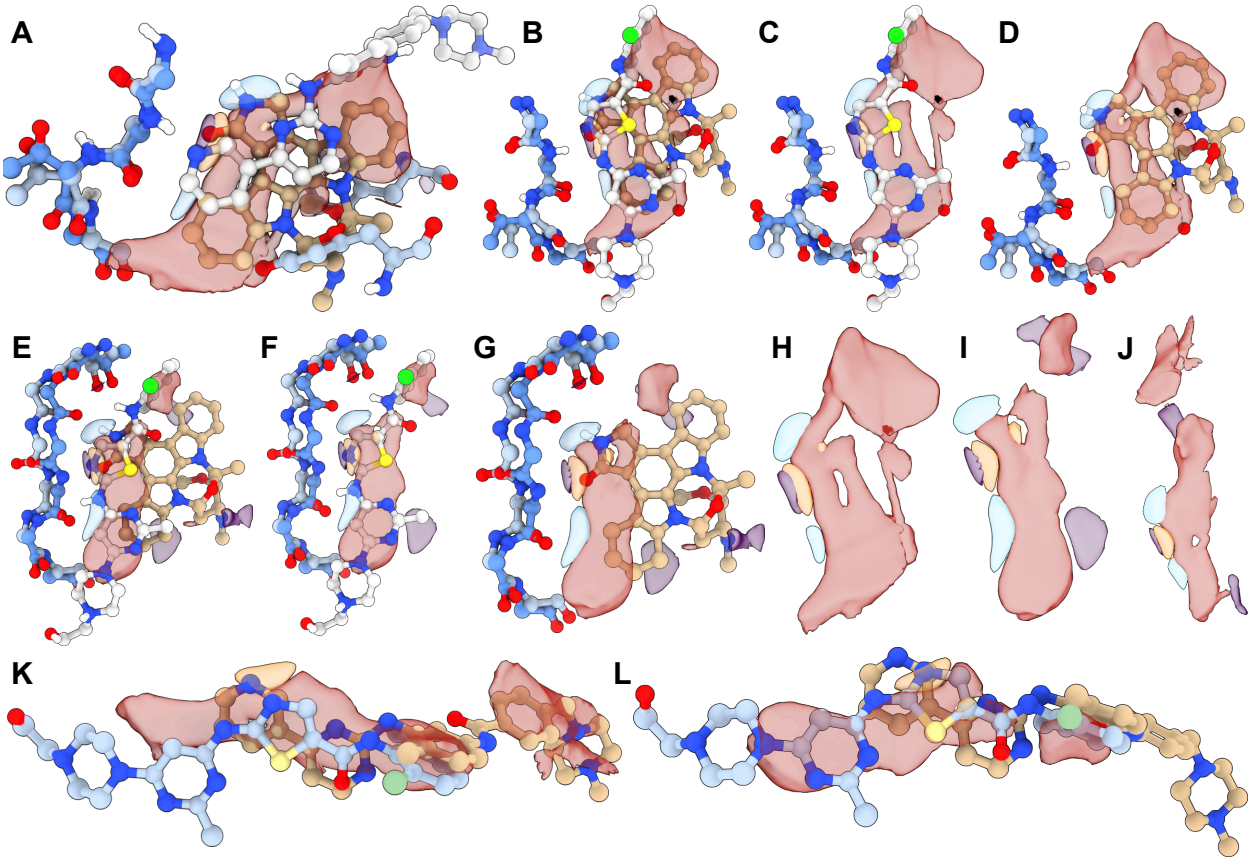


Figure S2: Ligand imprint in experimental crystal structures. A) Overlay of the crystal structure of imatinib in BCR-ABL (PDB-ID 2HYY, ligand color code white, protein color code light blue) with that of staurosporine in BCR-ABL (PDB-ID 2HZ4, ligand color code brown, protein color code darker blue). TPMs predicted for the latter structure reflect the ligand's chemical composition in the pocket and differ from the TPM predictions for imatinib in BCR-ABL (see Figure 2). Similar overlay with dasatinib (PDB-ID 2GQG) instead of imatinib. Dasatinib inherits imatinib's color code in this block. C) Overlay of dasatinib's structure with the TPM predictions for the crystal 2HZ4, which contains staurosporine in BCR-ABL. D) Staurosporine overlaid with the TPMs for 2HZ4. E) Overlay of dasatinib and staurosporine with the TPMs predicted for the crystal structure 2GQG, which contains dasatinib in BCR-ABL. Color code identical to that of block B). F) Dasatinib overlaid with the TPMs predicted for 2GQG, the crystal structure of dasatinib in BCR-ABL. G) Staurosporine overlaid with the TPMs predicted for 2GQG. H) TPMs for 2HZ4 (staurosporine in BCR-ABL). The latter are to be compared against predictions for I) 2GQG (dasatinib in BCR-ABL), and J) 2HYY (imatinib in BCR-ABL). K) Overlay of imatinib (brown) and dasatinib (light blue) with the TPMs for the PDB 2HYY (imatinib in BCR-ABL). L) Overlay of imatinib (brown) and dasatinib (light blue) with the TPMs for the PDB 2GQG (dasatinib in BCR-ABL).

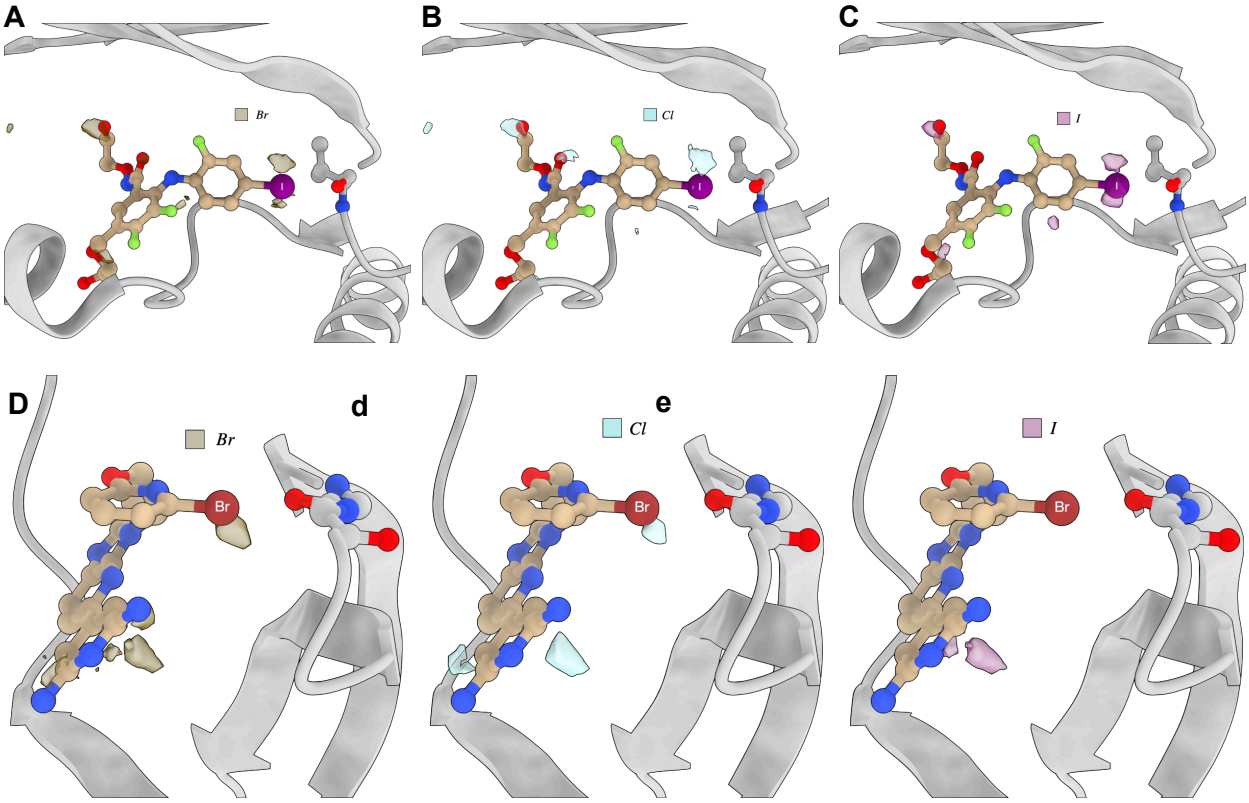


Figure S3. The specificity of halogen signals in TPM predictions. A-C) TPM predictions for the crystal structure 3DV3, the structure of the MEK kinase with halogen-containing inhibitors. A) The bromine, B) chlorine, and C) iodine predictions agree to a good extent, although some signals are specific to iodine atoms. D-F) TPM predictions for the crystal structure 7AKB, containing DYRK1A with a halogen-containing ligand. Predictions for D) bromine and E) chlorine are almost interchangeable, but the same does not apply to F) iodine.

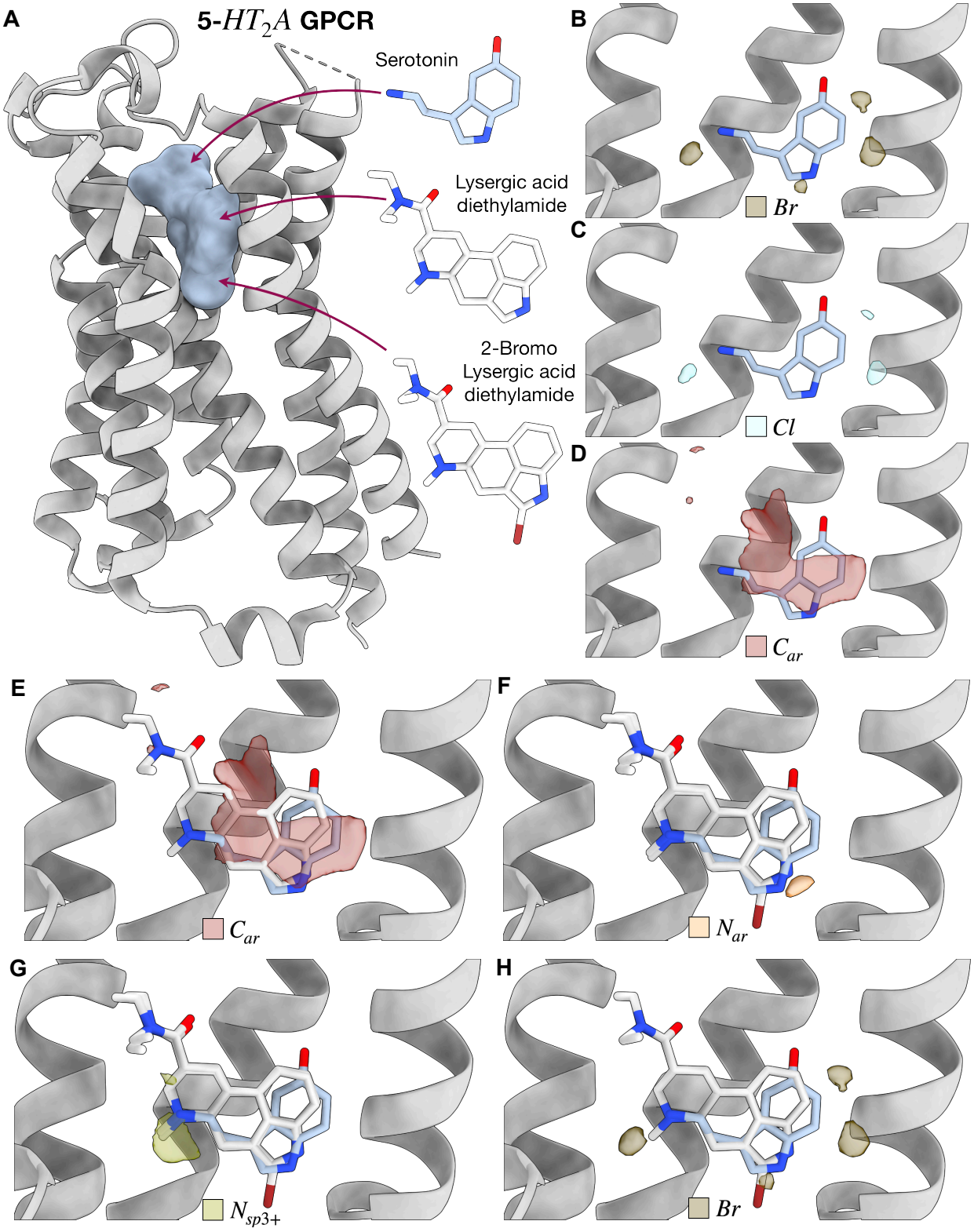
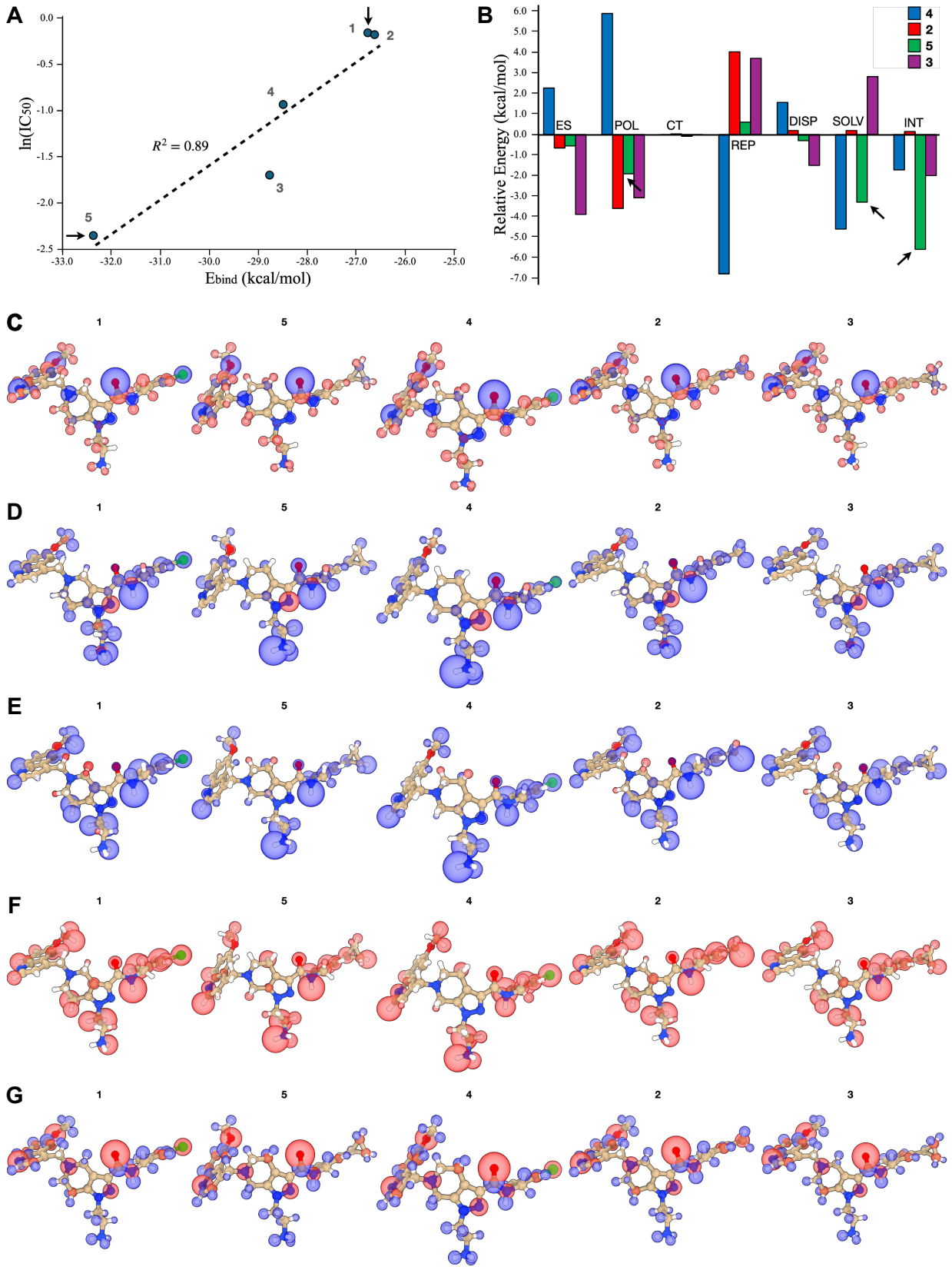


Figure S4. TPM predictions for the 5-HT2A GPCR. A) The receptor's structure with the ligand's density, which might be serotonin, lysergic acid diethylamide, or its 2-bromo derivative. B-D) Overlay of B) bromine, C) chlorine, and D) aromatic carbon densities predicted from the cryoEM structure of the receptor with serotonin. E) Overlay of the above-mentioned predicted aromatic carbon clouds and the structures of serotonin and lysergic acid diethylamide.

Similar overlays, but we consider F) aromatic nitrogens or G) ammonium nitrogens instead. H) Overlay of serotonin, 2-bromo lysergic acid diethyl amide and the bromine signals predicted for the GPCR with serotonin in the pocket.


A
ln(IC50)
Ebind (kcal/mol)
$R^2 = 0.89$
1
2
3
4
5
B
Relative Energy (kcal/mol)
ES
POL
CT
REP
DISP
SOLV
INT
4
2
5
3
C
D
E
F
G

511

Figure S5. Quantum Mechanical validation of the TPM-generated ligands. A) Correlation plot between the experimentally measured IC50 and the quantum mechanical binding energy. B) The relative Energy Decomposition and Deconvolution Analysis (EDDA) plot, which shows that compound **5**'s improved efficacy results from a mixture of improved polarization and solvation effects. C) Electrostatics maps for the 5 ligands analyzed. All maps are similar, indicating small fine differences. D) The total interaction maps for the ligand series, followed by E) the polarization maps, F) overlap-repulsion, and G) solvation.

518 **References**

519 1. Buttenschoen, M., Morris, G.M., and Deane, C.M. (2024). PoseBusters: AI-based docking methods fail to
520 generate physically valid poses or generalise to novel sequences. Chem. Sci. *15*, 3130–3139.

521 2. Škrinjar, P., Eberhardt, J., Durairaj, J., and Schwede, T. (2025). Have protein-ligand co-folding methods moved
522 beyond memorisation? bioRxiv. https://doi.org/10.1101/2025.02.03.636309.

523 3. Liu, F., Mailhot, O., Glenn, I.S., Vigneron, S.F., Bassim, V., Xu, X., Fonseca-Valencia, K., Smith, M.S.,
524 Radchenko, D.S., Fraser, J.S., et al. (2025). The impact of library size and scale of testing on virtual screening.
525 Nat. Chem. Biol. *21*, 1039–1045.

526 4. Hall, B.W., Tummino, T.A., Tang, K., Mailhot, O., Castanon, M., Irwin, J.J., and Shoichet, B.K. (2025). A
527 database for large-scale docking and experimental results. J. Chem. Inf. Model. *65*, 4458–4467.

528 5. Ghiandoni, G.M., Evertsson, E., Riley, D.J., Tyrchan, C., and Rathi, P.C. (2024). Augmenting DMTA using
529 predictive AI modelling at AstraZeneca. Drug Discov. Today *29*, 103945.

530 6. Edwards, A.M., Owen, D.R., and Structural Genomics Consortium Target 2035 Working Group (2025). Protein-
531 ligand data at scale to support machine learning. Nat. Rev. Chem. https://doi.org/10.1038/s41570-025-00737-z.

532 7. Chen, B., Huang, K., Raghupathi, S., Chandratreya, I., Du, Q., and Lipson, H. (2022). Automated discovery of
533 fundamental variables hidden in experimental data. Nat. Comput. Sci. *2*, 433–442.

534 8. Smith, J.S., Isayev, O., and Roitberg, A.E. (2017). ANI-1: an extensible neural network potential with DFT
535 accuracy at force field computational cost. Chem. Sci. *8*, 3192–3203.

536 9. Siebenmorgen, T., Menezes, F., Benassou, S., Merdivan, E., Didi, K., Mourão, A.S.D., Kitel, R., Liò, P.,
537 Kesselheim, S., Piraud, M., et al. (2024). MISATO: machine learning dataset of protein-ligand complexes for
538 structure-based drug discovery. Nat. Comput. Sci. *4*, 367–378.

539 10. Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2015). PDB-wide collection of
540 binding data: current status of the PDBbind database. Bioinformatics *31*, 405–412.

541 11. Zaucha, J., Softley, C.A., Sattler, M., Frishman, D., and Popowicz, G.M. (2020). Deep learning model predicts
542 water interaction sites on the surface of proteins using limited-resolution data. Chem. Commun. (Camb.) *56*,
543 15454–15457.

544 12. Durairaj, J., Adeshina, Y., Cao, Z., Zhang, X., Oleinikovas, V., Duignan, T., McClure, Z., Robin, X., Studer, G.,
545 Kovtun, D., et al. (2024). PLINDER: The protein-ligand interactions dataset and evaluation resource. bioRxiv.
546 https://doi.org/10.1101/2024.07.17.603955.

547 13. Grygier, P., Pustelny, K., Nowak, J., Golik, P., Popowicz, G.M., Plettenburg, O., Dubin, G., Menezes, F., and
548 Czarna, A. (2023). Silmitasertib (CX-4945), a clinically used CK2-kinase inhibitor with additional effects on
549 GSK3β and DYRK1A kinases: A structural perspective. J. Med. Chem. *66*, 4009–4024.

550 14. Dawidowski, M., Kalel, V.C., Napolitano, V., Fino, R., Schorpp, K., Emmanouilidis, L., Lenhart, D., Ostertag,
551 M., Kaiser, M., Kolonko, M., et al. (2020). Structure-activity relationship in pyrazolo[4,3-c]pyridines, first
552 inhibitors of PEX14-PEX5 protein-protein interaction with trypanocidal activity. J. Med. Chem. *63*, 847–879.

553 15. Dawidowski, M., Emmanouilidis, L., Kalel, V.C., Tripsianes, K., Schorpp, K., Hadian, K., Kaiser, M., Mäser, P.,
554 Kolonko, M., Tanghe, S., et al. (2017). Inhibitors of PEX14 disrupt protein import into glycosomes and kill
555 Trypanosoma parasites. Science *355*, 1416–1420.

556 16. Wahida, A., Buschhorn, L., Fröhling, S., Jost, P.J., Schneeweiss, A., Lichter, P., and Kurzrock, R. (2023). The
557 coming decade in precision oncology: six riddles. Nat. Rev. Cancer *23*, 43–54.

17. O'Boyle, N.M., Banck, M., James, C.A., Morley, C., Vandermeersch, T., and Hutchison, G.R. (2011). Open Babel: An open chemical toolbox. J. Cheminform. *3*, 33.

18. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152.

19. Menezes, F., and Popowicz, G.M. (2022). ULYSSES: An efficient and easy to use semiempirical library for C+. J. Chem. Inf. Model. *62*, 3685–3694.

20. Góral, I., Wichur, T., Sługocka, E., Grygier, P., Głuch-Lutwin, M., Mordyl, B., Honkisz-Orzechowska, E., Szałaj, N., Godyń, J., Panek, D., et al. (2024). Exploring novel GSK-3β inhibitors for anti-neuroinflammatory and neuroprotective effects: Synthesis, crystallography, computational analysis, and biological evaluation. ACS Chem. Neurosci. *15*, 3181–3201.

21. Nowacki, M., Menezes, F., Pykacz, E., Popiołek, M., Napolitano, V., Krishna, C.K., Kalel, V.C., Erdmann, R., Fröhlich, T., Plettenburg, O., et al. (2025). Quantum mechanics-driven structure-activity relationship study of PEX5-PEX14 protein-protein interaction inhibitors based on a dibenzo[b,e]azepin-6(6H)-one scaffold. Eur. J. Med. Chem., 117979.