A universal model for drug-receptor interactions

Filipe **Menezes**,[1,2,#,*] Adam **Wahida**,[3,#,*] Tony **Fröhlich**,[1,2,*] Phillip **Grass**,[4] Jan **Zaucha**,[1,2] Agata **Barzowska-Gogola**,[6,7] Anna **Czarna**,[7] Andreas **Hochhaus**,[8] Johannes **Nissen-Meyer**,[4] Marcus **Conrad**,[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.*

[6]*Łukasiewicz Research Network, Kraków Institute of Technology, 30-418 Kraków, Poland.*

[7]*Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland.*

[8]*Klinik für Innere Medizin II, Universitätsklinikum Jena, Jena, Germany.*

*these authors contributed equally: F. M., A. W., T. F.

§These authors co-supervised this study: M. C., R. K., 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)

Abstract

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 via non-bonded interactions. 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. Our understanding of, and reasoning from, non-bonded interaction chemistry is still too limited. Compounding this is the scarcity of novel chemistry and infinitesimal coverage of the chemical combinatorial space by current experimental data. To alleviate these problems, we show that a machine learning model can successfully learn and infer the principles of non-bonded interactions in the drug-receptor space. A reductionist approach to training data led to a model generalizing drug-target interactions to truly novel chemical matter without suffering from memorization bias. This work addresses that gap in drug discovery 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 and mutual adaptation of the binding molecules often render any physics-based models inadequate for capturing biological or clinical outcomes effectively. Current machine learning (ML) models treat these problems implicitly, relying on a latent space to find an approximate solution. Yet, with the current experimental data covering only an infinitesimal fraction of the possible combinatorial space, their predictive power diminishes sharply when applied to structurally diverse complexes.[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 historical chemical spaces, 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 experimental 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 drug's binding pose and affinity, increasing the DMTA turnover. Simultaneously transitioning from a historically biased, IP-crowded chemical combinatorial space to a novel, IP-free space is necessary, which largely remains unexplored.[6]

Here, we propose a theory-driven 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. We show that our model successfully learns non-bonded interaction chemistry principles and can infer drug chemical features. The model also proposes modifications that go beyond traditional understanding yet are experimentally proven to be correct.

A machine learning model to generalize non-bonded chemistry

We built 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, leads to drugs with 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,[9] is explored by only *ca.* 50,000 published protein-ligand complex structures.[10] 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 structural and bonding properties (*e.g.*, sp^3; **Figure 1C**), and their surrounding receptor amino acids constitute a microenvironment that corresponds to 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**).

In consequence, training becomes agnostic to drug chemistry and receptor identity, 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[11,12] used for training yield over 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 drug-ligand, satisfying all possible interactions the receptor could request. No real molecule can meet all, often exclusive, features: *e.g.*, aromatic nitrogens and sp^2 oxygens are mutually exclusive and incompatible atom types, which, however, are involved in hydrogen bonding. 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 real-life, wet lab drug optimization

problems. **Figure 2** shows TPM predictions for several receptor classes of biological targets, showcasing the generality of our approach.

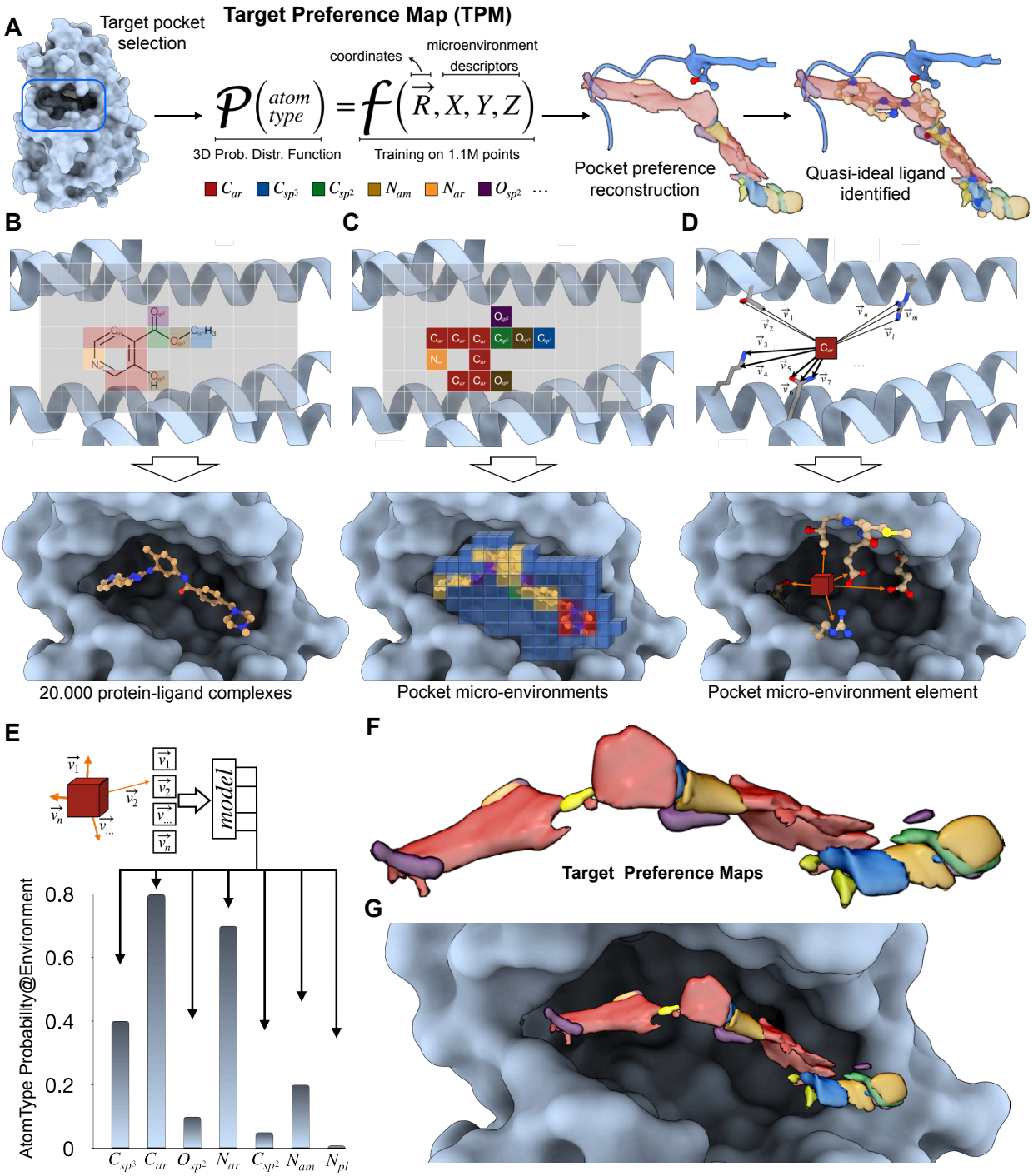


Figure 1: The Target Preference Map (TPM) Model. A) Conceptualization of the TPM model. Given the pocket of a biological target, we define atom-type-specific probability distribution functions. The function takes as arguments Cartesian coordinates and 3D chemical microenvironments (surrounding receptor atoms) 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 (TPMs), G) we generate target-specific localized preference functions.

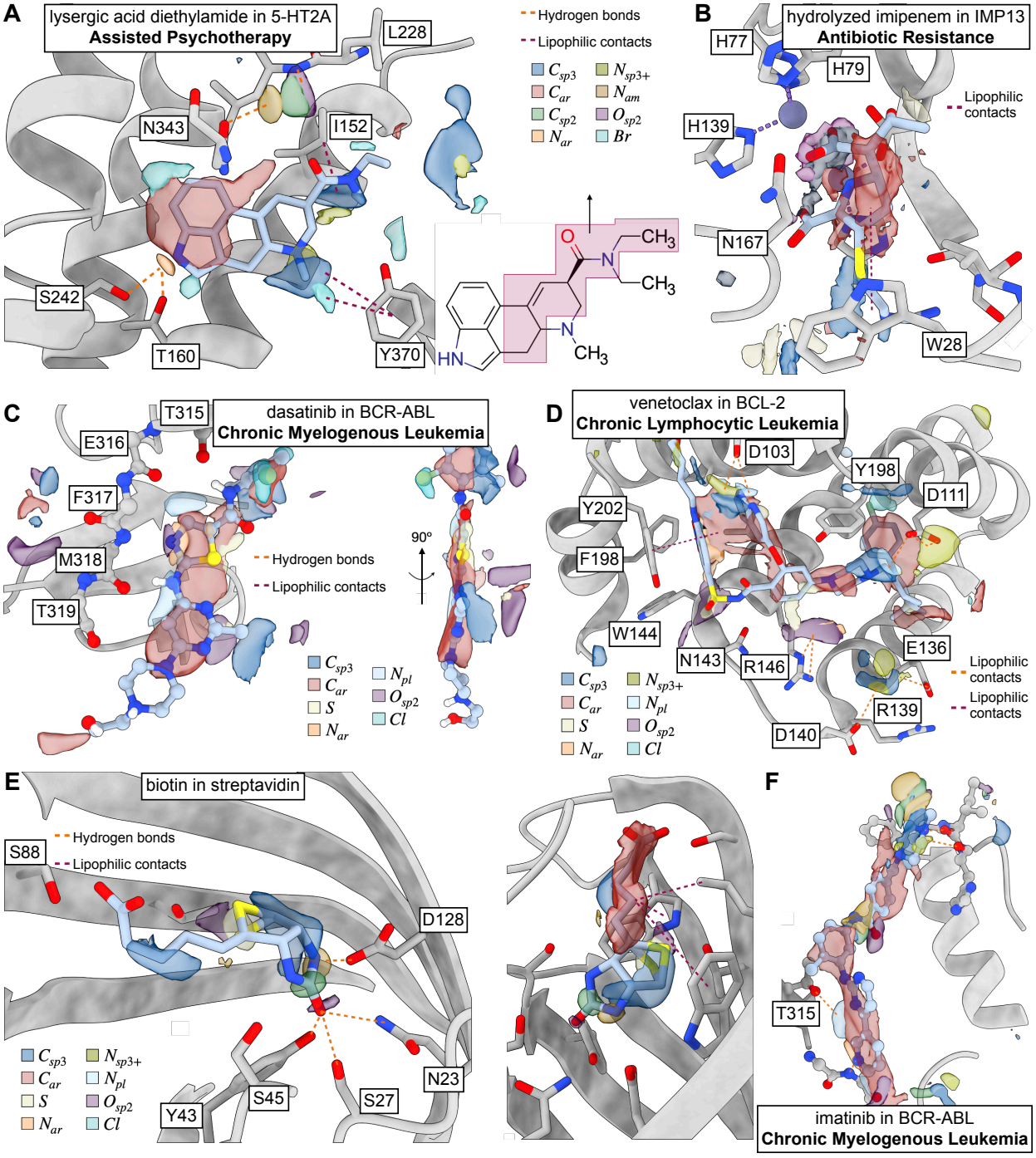


Figure 2. Examples of TPM correctly inferring drug features from diverse classes of receptors. A. Lysergic acid diethyl amide in the 5-HT2A G-Protein Coupled Receptor (GPCR), a model for assisted psychotherapy. TPM clouds predicted for the structure of the 5-HT2A GPCR with serotonin inside (PDB-ID 9ARX).[25] 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 sp^3 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 6RZR).[16] 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 Zn^{2+} cations as seen in known inhibitors. C) A type I kinase inhibitor, dasatinib, in its biological target BCR::ABL1 (PDB-ID 2GQG),[26] 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 sp^2 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),[27] 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 2F01).[28] The model predicts adding aliphatic carbons, amide groups, and sulfur atoms in good agreement with the ligand. F) TPM predictions for imatinib in BCR::ABL1's pocket (PDB-ID 2HYY),[29] 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.

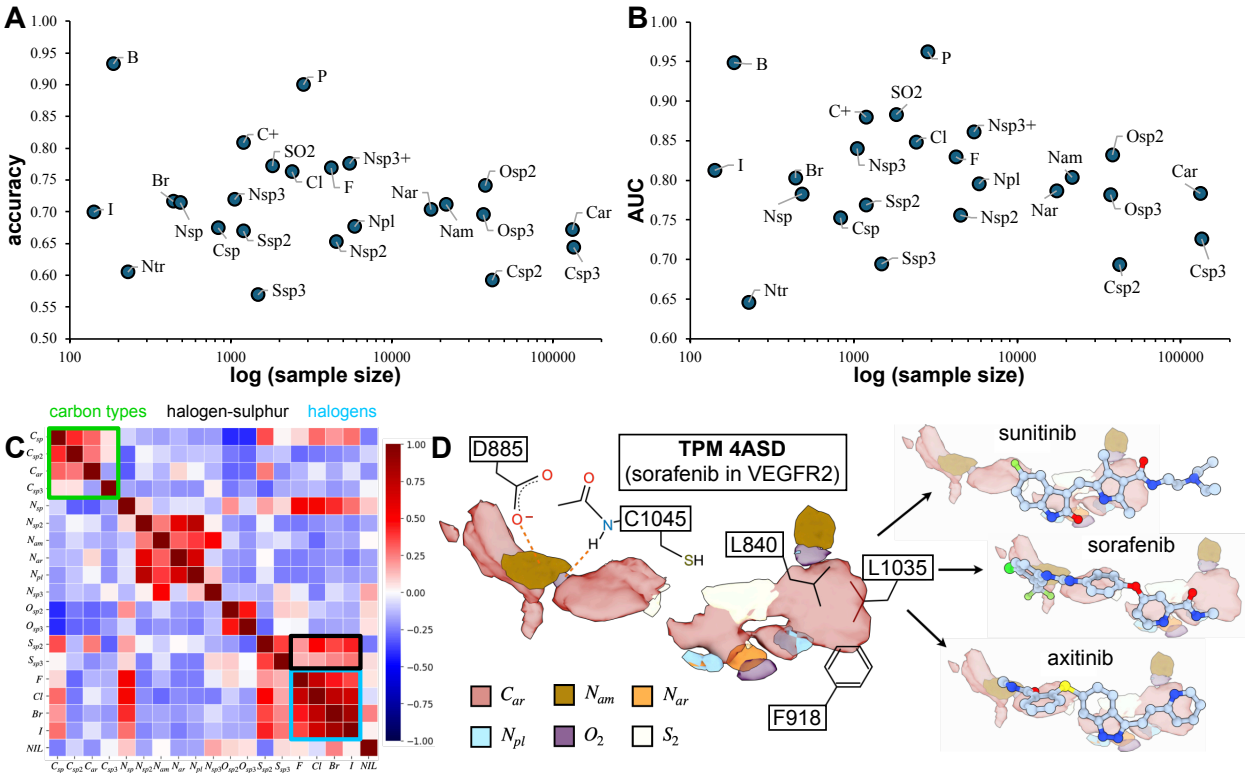


Figure 3. Statistics and Inference of the TPM model. A. TPM-model accuracy for every atom type considered against the respective sample size on a logarithmic scale. B. TPM-model AUC for every atom type considered against the respective sample size on a logarithmic scale. These two blocks show that polar 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? Observed (anti)correlations align perfectly with traditional medicinal chemistry substitutions. 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.

In our study, we consider the following atom classes: B–boron; C_X–carbon, where X takes the values *sp*, sp^2, and sp^3 for specific atomic hybridizations and *ar* for aromatic carbons; N_X–nitrogen, where X takes similar values as carbon types, with the additional classes *am* for amide nitrogens and *pl* for non-amide planar nitrogens; O_X–oxygen, with X taking values similar to carbon and nitrogen hybridization states; P–phosphorous; S_X–sulphur, where 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 drug-like-molecule-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 type metrics are mostly independent of sample size (**Figures 3A**, **B**). The majority of atom types show good accuracy (>70%), with phosphorus and boron reaching excellent values (>90%), the latter nearing 95%. Such high accuracy may suggest overfitting due to a small sample size (~200 samples) and is likely resulting from the nearly identical structural properties required to observe binding of boron atoms to a protein–*e.g.*, hydrolytic enzymes active sites are frequently targeted by boronic acid warheads. 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 types with limited accuracy despite a large 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 (–0.5 electrons vs. –0.25 electrons), 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 clearly defined, directional hydrogen bonding, while C_{ar} favors less directional, lipophilic contacts and π-stacking, explaining the model's lower accuracy metrics. 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 predominantly by chemical interaction types rather than by data quantity: atom types involved in polar and hydrogen bonding tend to be predicted more accurately. Conversely, atom types involved in lipophilic interactions are where the model shows 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 very nature of chemical interactions.

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 strongly. 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 the functional group C≡N 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, nor the directionality of the resulting hydrogen bonds. Oxygen predictions also show a 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 to the protein.[13]

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 participate in hydrogen bond networks, 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 interactions. 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 drug 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 respective structural family from training. We took the structures of VEGFR2 with sorafenib, axitinib, and sunitinib as reference (respectively, PDB-IDs 4ASD, 4AGC, and 4AGD; **Figure 3D**).[14] 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 inhibitors, axitinib and sunitinib. Sunitinib shows the least overlap with the 4ASD TPM set. Initially, we attributed this to the fact that sunitinib (like axitinib) is a type I kinase inhibitor, while sorafenib is a type II inhibitor. However, superimposing the three PDB structures reveals a striking similarity in the VEGFR conformational states. This suggests that the differences arise from the distinct ways each inhibitor interacts with the kinase's ATP-binding pocket.

Several structural elements are nonetheless conserved, and the TPM model suggests how to explore the receptor-drug interactions best. This is, for instance, the alignment of an N_{pl} signal with sunitinib's pyrrole ring, along with the oxoindole part of the drug. 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 drug's amide group, and the drug 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 drug not accessing that area of the pocket. This signal, which cannot stem from drug 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 drug'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 drug 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. However, and expectedly, the TPMs for type I and II inhibitors differ noticeably. 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 (accession codes 7Z5N, 7Z1F, and 7Z1G; **Figure 4**).[15] To minimize the possible drug-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 their position in the pocket, 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.

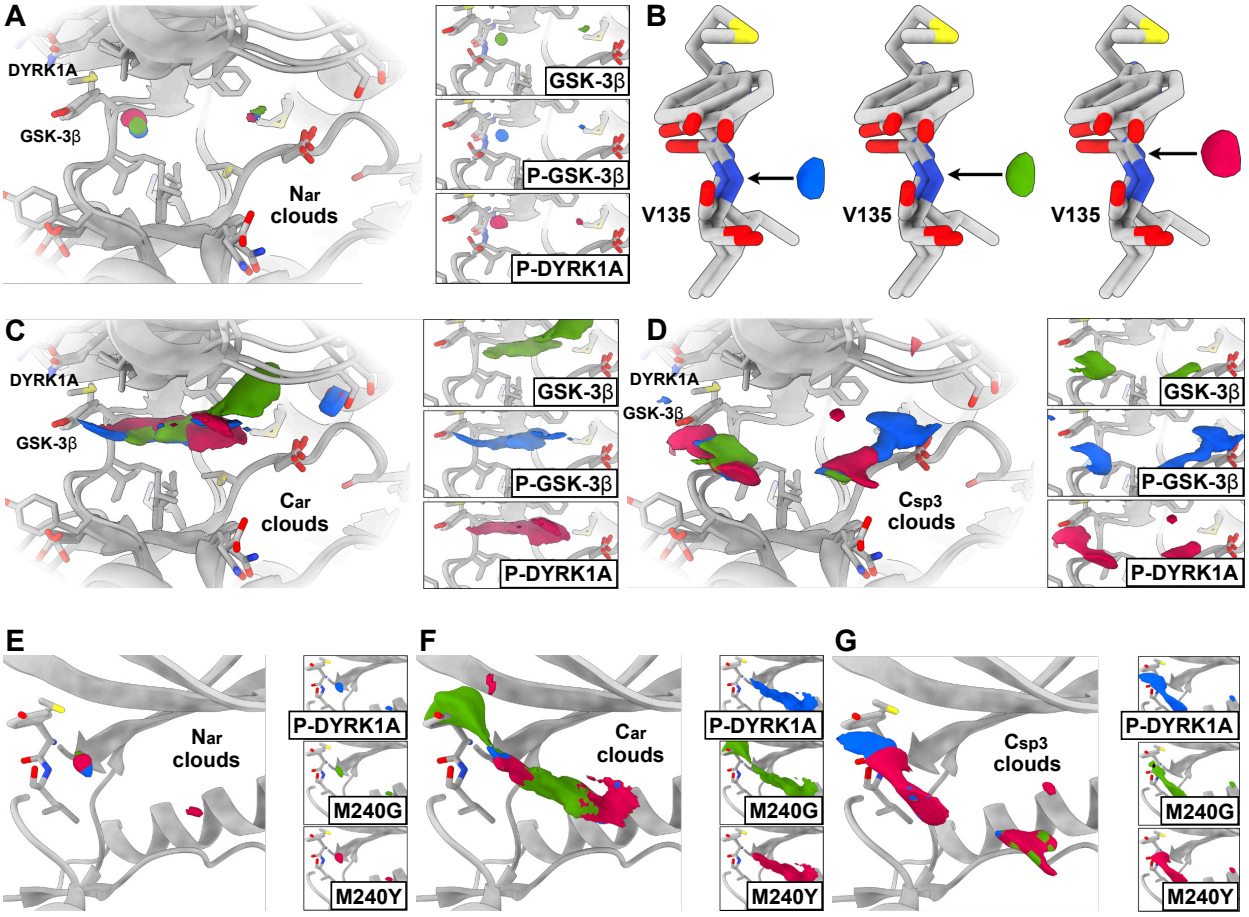


Figure 4: The TPM model understands specific non-bonded chemistry despite target similarity. In this analysis, we looked at TPM predictions for GSK3β and DYRK1A. A) TPMs for aromatic nitrogens follow the coordinates of hinge residues in GSK3β (phosphorylated and unphosphorylated, PDB-IDs 7Z1G and 7Z1F) and phosphorylated DYRK1A (PDB-ID 7Z5N).[15] 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.

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.[16] The IMP-13 coordinates two adjacent zinc ions at its active site.

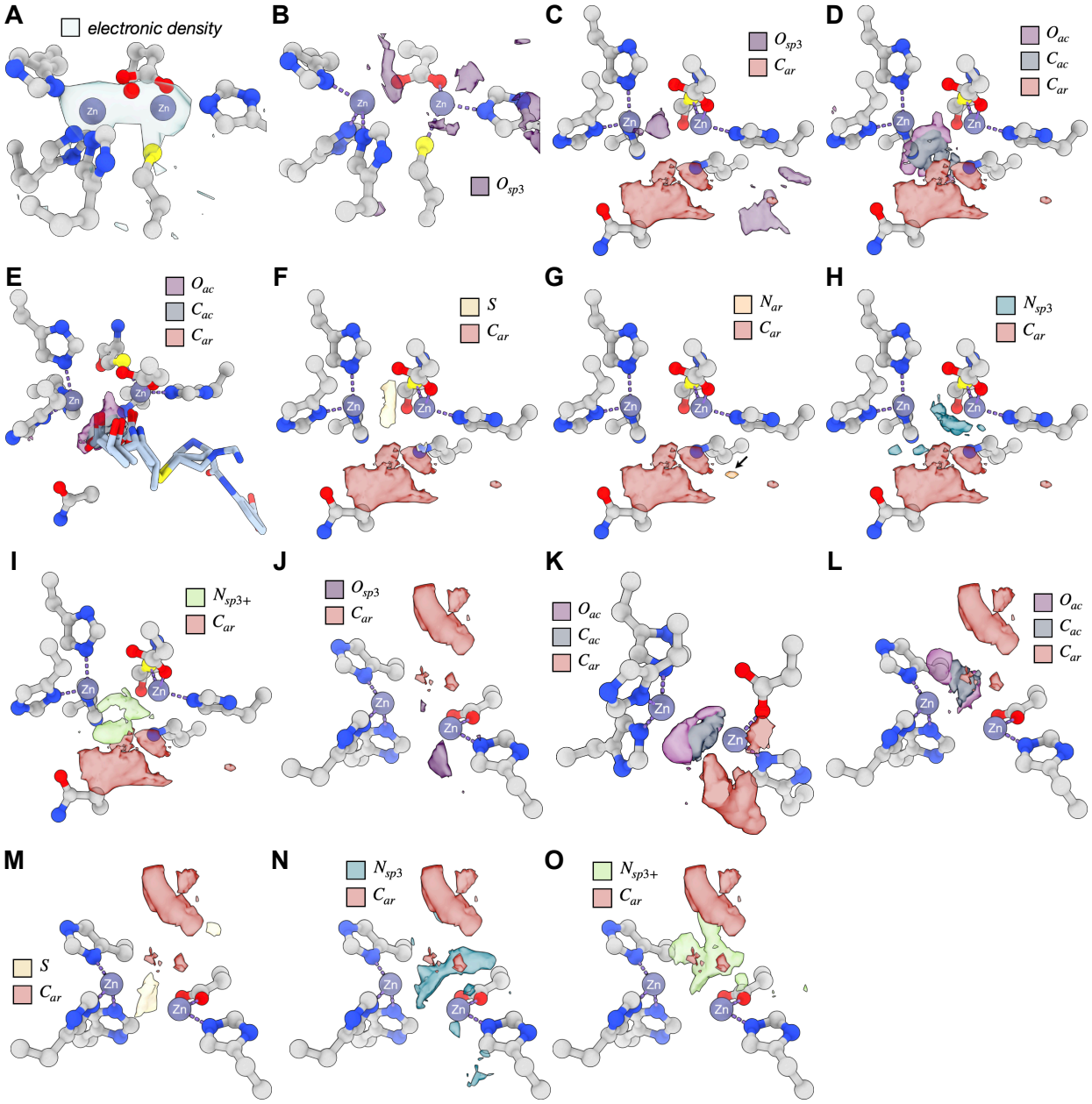


Figure 5. TPM predictions at the quantum chemistry level for the IMP13 metalloenzyme. A) The experimental electronic density of apo IMP13 (PDB-ID 6R78) showing the presence of an oxygen atom between the zinc cations (hydroxy group).[16] B-I) TPM predictions for IMP13 with hydrolysed meropenem (PDB-ID 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 rationalized by classical physics only by quantum chemistry. 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 known drug-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 drug 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 placements in binding pockets, guide towards the development of 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 in previous attempts (**Figure 6a)**.[17,18] Classical optimization reached an apparent plateau: conventional structure-based drug design strategies and traditional medicinal chemistry approaches failed to yield further improvements in the compounds' 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 modifications (**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.[17,18] 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: the selectivity index (SI) increased to 42.1, against the previous value of 13.4 (SI = EC_{50} (*T. cruzi*)/CC_{50} in 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.

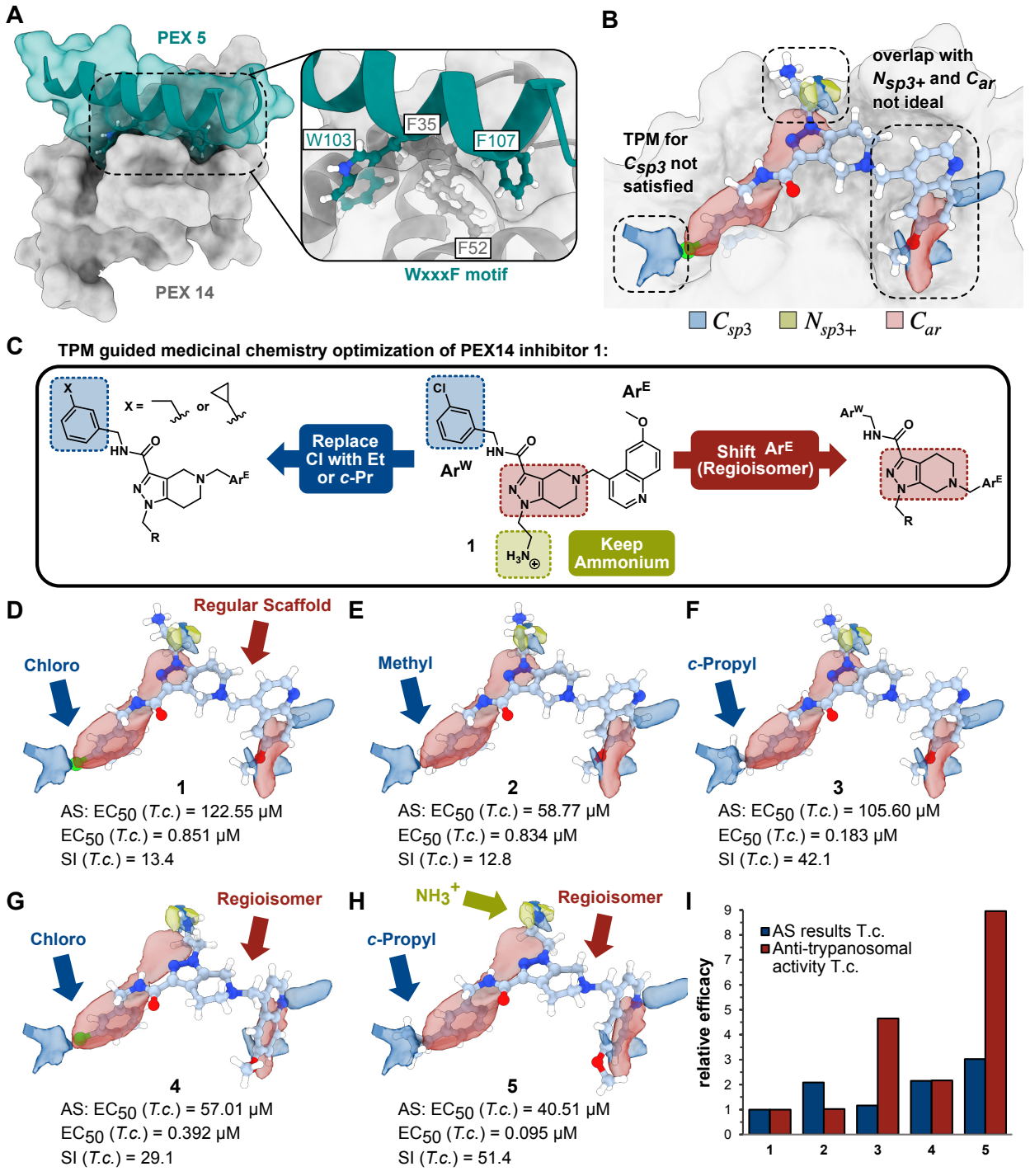


Figure 6: Prospective validation of the TPM model on a challenging protein-protein interaction (PPI) target. A) The PPI interface of PEX14 and PEX5, and the main residues involved in the interaction. B) TPM predictions based on the 5L87 crystal structure, containing PEX14 bound to one inhibitor.[17] The maps reveal three possible structural modifications: i) introduction of alkyl groups on the leftmost aromatic substituent (also termed western aromatic moiety); ii) shift of the rightmost aromatic group (also termed eastern aromatic moiety) to match the aromatic density prediction; iii) and introducing an ethyl ammonium group in the center of the ligand to capture the negatively charged residues Glu34 and Asp38. 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-H) The structure and

327 efficacy of several inhibitors, compounds **1**-**5**, along with their target selectivity. I) The relative efficacy of all
328 compounds herein studied.

329 Modifying the core scaffold of a lead compound, a strategy known as scaffold-hopping, is often synthetically
330 demanding, as it requires substantial alterations of synthetic routes. Therefore, constructive guidance is essential
331 before the effort is made. Following the TPMs, we generated regioisomers that exhibited a consistent increase in
332 activity. The regioisomer **4** demonstrated a twofold enhancement in pharmacological potency relative to inhibitor **1**,
333 accompanied by superior alignment with the generated TPMs. This conformational adjustment also optimizes the
334 ligand's ammonium group placement, which should better overlap with the predicted N_{sp3+} density, situated between
335 the two negatively charged residues. Intuitively, an ammonium group between Glu34 and Asp38 facilitates hydrogen
336 bonds and salt bridges. Regioisomerisation is also expected to improve compound **4**'s quinoline moiety (Ar^E)
337 alignment with the C_{ar} TPM compared to derivatives containing the regular pyrazolo[4,3-*c*]pyridine scaffold
338 (compounds **1**-**3**). Finally, the target molecule **5**, which incorporates all previously established modifications and
339 exhibits the best overlap with the generated TPMs, emerged as the most potent inhibitor across all biological assays,
340 enhancing trypanocidal activity by nearly tenfold, compared to the previous lead compound **1**. These findings highlight
341 the exceptional predictive accuracy of the TPM model, which is an essential feature for drug design, often lacking in
342 current state-of-the-art *in silico* tools. Moreover, this result exemplifies why such a modification would likely be
343 overlooked in conventional lead optimization campaigns. This result highlights the practical value of the TPM
344 approach in drug design, mainly as it directs researchers toward unconventional and subtle medicinal chemistry
345 modifications that will otherwise remain unexplored. We also performed a quantum mechanical evaluation of the new
346 series of inhibitors to identify a physically motivated rationale behind the improved efficacy. The results, detailed in
347 Figure S5, show an excellent correlation between TPM-suggestions, experimental affinity, and the quantum
348 mechanical evaluations, providing yet additional validation of the model.

349 **Discussion**

350 We have demonstrated that an ML model can correctly learn and infer principles of non-bonded interaction chemistry
351 from experimental structural data. The TPM mapping provides a probabilistic model of an ideal drug for each receptor
352 structure. The model's predictions extend beyond classical physics and chemistry and can be only fully rationalized
353 by quantum chemistry interpretation. Thus, we show that, by improved treatment of experimental data we can train a
354 model to infer not only classical, but also quantum interpretation of molecules. The model's approach removes drug-
355 receptor interactions from latent space towards an explainable, physical model that can be used directly for drug
356 optimization, or, in the future, to dock and generate new molecules. The model's sensitivity to minute structural detail
357 allows calculation of drug preference changes at minimal receptor alterations like point mutations.

358 On the general level, advances in computational biology have revealed the predictive power of genomic data,
359 especially when coupled with machine learning models, enabling a shift from serendipitous discovery toward
360 systematic protein structure generation. In principle, the genomic blueprint of a disease harbors the information needed
361 to guide rational drug development. This, when coupled with a detailed understanding of the molecular underpinnings
362 of disease, may deliver tailored therapies, realizing the envisioned goal of precision medicine. Yet, current strategies,
363 such as inferring structure from sequence, fall short of capturing the atomistic detail required for therapeutic design.
364 A key bottleneck remains in constructing models that can learn the rules of drug-target recognition. To this end, we
365 present a ML platform for small-molecule discovery, trained on over 1.1 million drug-protein interactions. This
366 framework identifies binding determinants and optimizes drug-like properties with atomic resolution. Our integrated
367 approach recapitulates the chemical features of existing therapeutics, enhances suboptimal leads, and accelerates *de*
368 *novo* design against challenging targets, all with minimal design latency.

369 The complex nature of biological structural data and the limited diversity of training sets hinder the training of robust
370 and reliable ML models. Complex dynamic, or quantum mechanical problems like drug-receptor binding, are
371 delegated to the latent space. Compared with the nearly limitless chemical diversity, the experimental noise and the
372 limits of training data render these models susceptible to overfitting and poor generalizability. Even the best-
373 performing models suffer from heavy memorization and dramatically lose accuracy when dealing with molecules
374 deviating from their training sets.[1,2] We shifted the focus from entire molecules to local atomic microenvironments to
375 mitigate this. Thus, we divided the available drug-biomolecule data into local structural features that are chemically
376 diverse and numerous to allow the model to generalize non-bonded drug-receptor interactions. This renders the TPM
377 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 TPM point carries chemical knowledge of all structures analyzed instead of relying on a small set of force-field-like 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 drug'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.[11] TPMs can be directly interpreted by humans or used in structure-aware software for docking or *de novo* drug 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 necessitates a flexible tool to connect genetic information with drug chemistry interactions,[19] 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 TPMs offer a novel framework for generalizing non-bonded interactions, several limitations merit consideration. 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. Future models can be trained on molecular dynamics traces to improve this situation.[11] Moreover, 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. 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. This will allow to finally solve the structure-based drug discovery limitations.

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,[20] ensuring consistency with the input featurization of protein atoms. To ensure rotational invariance, we define a local coordinate system at each query point.[13] 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.[12] 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.[21] 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 (**Figure 6D**). 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** (**Figure 6D**), 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 one of the ammonium group of compound **1** did not achieve optimal alignment with the predicted N_{sp3+} density maps. We hypothesized that this alignment could improve for the planned regioisomer derivatives.

A detailed description of the synthetic routes is available in supplement.

Quantum Mechanical calculations. To further validate the TPM-suggested ligand modifications we ran quantum mechanical evaluations using our in-house ULYSSES software.[22] Validation and analysis was performed using quantum mechanical analysis protocols as described in previous work.[23,24]

Code Availability. The source code for the TPM model will be released on GitHub following manuscript approval. Additionally, a web server for TPM generation and TPM-based docking will also be made available at that time.

Acknowledgements

This work was supported by the Bundesministerium für Bildung und Forschung (BMBF), projects SUPREME (number: 031L0268) and DATIPilot Sprint–MISATO (03DPS1234); by the Bundesministerium für Wirtschaft und Klimaschutz (BMWK) via ZIM (grant number: KK5710001BA4); by the Deutsche Forschungsgemeinschaft, DFG grant CRC387, project ID TRR 387/1-514894665.

Declaration of conflicts of interest

The use of the TPM technology for drug discovery is protected by a joint Helmholtz Munich and Khumbu.AI patent application (EP2023083578; Publication Number: 2024/115584). J. Z. is currently affiliated with Astra Zeneca.

References

(1) Buttenschoen, M.; Morris, G. M.; Deane, C. M. PoseBusters: AI-Based Docking Methods Fail to Generate Physically Valid Poses or Generalise to Novel Sequences. *Chem. Sci.* **2024**, *15* (9), 3130–3139. https://doi.org/10.1039/D3SC04185A.

(2) Škrinjar, P.; Eberhardt, J.; Tauriello, G.; Schwede, T.; Durairaj, J. Have Protein-Ligand Co-Folding Methods Moved beyond Memorisation? February 7, 2025. https://doi.org/10.1101/2025.02.03.636309.

(3) Liu, F.; Mailhot, O.; Glenn, I. S.; Vigneron, S. F.; Bassim, V.; Xu, X.; Fonseca-Valencia, K.; Smith, M. S.; Radchenko, D. S.; Fraser, J. S.; Moroz, Y. S.; Irwin, J. J.; Shoichet, B. K. The Impact of Library Size and Scale of Testing on Virtual Screening. *Nat. Chem. Biol.* **2025**, *21* (7), 1039–1045. https://doi.org/10.1038/s41589-024-01797-w.

(4) Hall, B. W.; Tummino, T. A.; Tang, K.; Mailhot, O.; Castanon, M.; Irwin, J. J.; Shoichet, B. K. A Database for Large-Scale Docking and Experimental Results. *J. Chem. Inf. Model.* **2025**, *65* (9), 4458–4467. https://doi.org/10.1021/acs.jcim.5c00394.

(5) Ghiandoni, G. M.; Evertsson, E.; Riley, D. J.; Tyrchan, C.; Rathi, P. C. Augmenting DMTA Using Predictive AI Modelling at AstraZeneca. *Drug Discov. Today* **2024**, *29* (4), 103945. https://doi.org/10.1016/j.drudis.2024.103945.

(6) Edwards, A. M.; Owen, D. R.; The Structural Genomics Consortium Target 2035 Working Group; Zhang, L.; Young, D. W.; Willson, T. M.; Wellnitz, J.; Wang, Y.; Walsh, J.; Vernet, E.; Tropsha, A.; Tredup, C.; Todd, M. H.; Tjaden, A.; Thamm, S.; Sundström, M.; Steffen, A.; Stauffer, S.; De Souza, L. R.; Shen, M.; Schütt, K.; Schiavone, L. H.; Schapira, M.; Santhakumar, S.; Saikatendu, K.; Rivers, E.; Petrović, D.; Peng, H.; O'Donnell, J. P.; Müller-Knapp, S.; Mueller-Fahrnow, A.; Morgan, M. R.; Montel, F.; Mobarec, J. C.; Michel, M.; Melliou, S.; Lessel, U.; Leach, A. R.; Krämer, O.; Krieger, F.; Knapp, S.; Keefe, A.; Kannt, A.; Johnson, S. A.; Häberle, S.; Holzinger, E. R.; Hartung, I. V.; Harding, R. J.; Hanke, T.; Halabelian, L.; Haibe-Kains, B.; Günther, J.; Guié, M.-A.; Gordijo, C.; Gileadi, O.; Foschini, L.; Fernández-Montalván, A.; Engkvist, O.; Edwards, M. M.; Duerr, K.; Drewry, D.; Dou, D.; Djordjevic, S.; Diaz, A. S.; Martinez Cuesta, S.; Counago, R.; Cornell, W. D.; Coker, J. A.; Clevert, D.-A.; Cernak, T.; Burgess-Brown, N. A.; Brown, P. J.; Bengtson, M. H.; Bashore, F. M.; Barsyte-Lovejoy, D.; Baghaie, A. J.; Axtman, A. D.; Arrowsmith, C.; Antolin, A. A.; Ackloo, S. Protein–Ligand Data at Scale to Support Machine Learning. *Nat. Rev. Chem.* **2025**. https://doi.org/10.1038/s41570-025-00737-z.

(7) Chen, B.; Huang, K.; Raghupathi, S.; Chandratreya, I.; Du, Q.; Lipson, H. Automated Discovery of Fundamental Variables Hidden in Experimental Data. *Nat. Comput. Sci.* **2022**, *2* (7), 433–442. https://doi.org/10.1038/s43588-022-00281-6.

(8) Smith, J. S.; Isayev, O.; Roitberg, A. E. ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost. *Chem. Sci.* **2017**, *8* (4), 3192–3203. https://doi.org/10.1039/C6SC05720A.

(9) Polishchuk, P. G.; Madzhidov, T. I.; Varnek, A. Estimation of the Size of Drug-like Chemical Space Based on GDB-17 Data. *J. Comput. Aided Mol. Des.* **2013**, *27* (8), 675–679. https://doi.org/10.1007/s10822-013-9672-4.

(10) Durairaj, J.; Adeshina, Y.; Cao, Z.; Zhang, X.; Oleinikovas, V.; Duignan, T.; McClure, Z.; Robin, X.; Studer, G.; Kovtun, D.; Rossi, E.; Zhou, G.; Veccham, S.; Isert, C.; Peng, Y.; Sundareson, P.; Akdel, M.; Corso, G.; Stärk, H.; Tauriello, G.; Carpenter, Z.; Bronstein, M.; Kucukbenli, E.; Schwede, T.; Naef, L. PLINDER: The Protein-Ligand Interactions Dataset and Evaluation Resource. July 17, 2024. https://doi.org/10.1101/2024.07.17.603955.

(11) Siebenmorgen, T.; Menezes, F.; Benassou, S.; Merdivan, E.; Didi, K.; Mourão, A. S. D.; Kitel, R.; Liò, P.; Kesselheim, S.; Piraud, M.; Theis, F. J.; Sattler, M.; Popowicz, G. M. MISATO: Machine Learning Dataset of Protein–Ligand Complexes for Structure-Based Drug Discovery. *Nat. Comput. Sci.* **2024**, *4* (5), 367–378. https://doi.org/10.1038/s43588-024-00627-2.

(12) Liu, Z.; Li, Y.; Han, L.; Li, J.; Liu, J.; Zhao, Z.; Nie, W.; Liu, Y.; Wang, R. PDB-Wide Collection of Binding Data: Current Status of the PDBbind Database. *Bioinformatics* **2015**, *31* (3), 405–412. https://doi.org/10.1093/bioinformatics/btu626.

(13) Zaucha, J.; Softley, C. A.; Sattler, M.; Frishman, D.; Popowicz, G. M. Deep Learning Model Predicts Water Interaction Sites on the Surface of Proteins Using Limited-Resolution Data. *Chem. Commun.* **2020**, *56* (98), 15454–15457. https://doi.org/10.1039/D0CC04383D.

(14) McTigue, M.; Murray, B. W.; Chen, J. H.; Deng, Y.-L.; Solowiej, J.; Kania, R. S. Molecular Conformations, Interactions, and Properties Associated with Drug Efficiency and Clinical Performance among VEGFR TK Inhibitors. *Proc. Natl. Acad. Sci.* **2012**, *109* (45), 18281–18289. https://doi.org/10.1073/pnas.1207759109.

(15) Grygier, P.; Pustelny, K.; Nowak, J.; Golik, P.; Popowicz, G. M.; Plettenburg, O.; Dubin, G.; Menezes, F.; Czarna, A. Silmitasertib (CX-4945), a Clinically Used CK2-Kinase Inhibitor with Additional Effects on GSK3β and DYRK1A Kinases: A Structural Perspective. *J. Med. Chem.* **2023**, *66* (6), 4009–4024. https://doi.org/10.1021/acs.jmedchem.2c01887.

(16) Softley, C. A.; Zak, K. M.; Bostock, M. J.; Fino, R.; Zhou, R. X.; Kolonko, M.; Mejdi-Nitiu, R.; Meyer, H.; Sattler, M.; Popowicz, G. M. Structure and Molecular Recognition Mechanism of IMP-13 Metallo-β-Lactamase. *Antimicrob. Agents Chemother.* **2020**, *64* (6), e00123-20. https://doi.org/10.1128/AAC.00123-20.

(17) Dawidowski, M.; Emmanouilidis, L.; Kalel, V. C.; Tripsianes, K.; Schorpp, K.; Hadian, K.; Kaiser, M.; Mäser, P.; Kolonko, M.; Tanghe, S.; Rodriguez, A.; Schliebs, W.; Erdmann, R.; Sattler, M.; Popowicz, G. M. Inhibitors of PEX14 Disrupt Protein Import into Glycosomes and Kill *Trypanosoma* Parasites. *Science* **2017**, *355* (6332), 1416–1420. https://doi.org/10.1126/science.aal1807.

(18) Dawidowski, M.; Kalel, V. C.; Napolitano, V.; Fino, R.; Schorpp, K.; Emmanouilidis, L.; Lenhart, D.; Ostertag, M.; Kaiser, M.; Kolonko, M.; Tippler, B.; Schliebs, W.; Dubin, G.; Mäser, P.; Tetko, I. V.; Hadian, K.; Plettenburg, O.; Erdmann, R.; Sattler, M.; Popowicz, G. M. Structure–Activity Relationship in Pyrazolo[4,3- *c*]Pyridines, First Inhibitors of PEX14–PEX5 Protein–Protein Interaction with Trypanocidal Activity. *J. Med. Chem.* **2020**, *63* (2), 847–879. https://doi.org/10.1021/acs.jmedchem.9b01876.

(19) Wahida, A.; Buschhorn, L.; Fröhling, S.; Jost, P. J.; Schneeweiss, A.; Lichter, P.; Kurzrock, R. The Coming Decade in Precision Oncology: Six Riddles. *Nat. Rev. Cancer* **2023**, *23* (1), 43–54. https://doi.org/10.1038/s41568-022-00529-3.

(20) O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R. Open Babel: An Open Chemical Toolbox. *J. Cheminformatics* **2011**, *3* (1), 33. https://doi.org/10.1186/1758-2946-3-33.

(21) Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W. CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data. *Bioinformatics* **2012**, *28* (23), 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.

(22) Menezes, F.; Popowicz, G. M. *ULYSSES* : An Efficient and Easy to Use Semiempirical Library for C++. *J. Chem. Inf. Model.* **2022**, *62* (16), 3685–3694. https://doi.org/10.1021/acs.jcim.2c00757.

(23) Góral, I.; Wichur, T.; Sługocka, E.; Grygier, P.; Głuch-Lutwin, M.; Mordyl, B.; Honkisz-Orzechowska, E.; Szałaj, N.; Godyń, J.; Panek, D.; Zaręba, P.; Sarka, A.; Żmudzki, P.; Latacz, G.; Pustelny, K.; Bucki, A.; Czarna, A.; Menezes, F.; Więckowska, A. Exploring Novel GSK-3β Inhibitors for Anti-Neuroinflammatory and Neuroprotective Effects: Synthesis, Crystallography, Computational Analysis, and Biological Evaluation. *ACS Chem. Neurosci.* **2024**, acschemneuro.4c00365. https://doi.org/10.1021/acschemneuro.4c00365.

(24) Nowacki, M.; Menezes, F.; Pykacz, E.; Popiołek, M.; Napolitano, V.; Krishna, C. K.; Kalel, V. C.; Erdmann, R.; Fröhlich, T.; Plettenburg, O.; Sattler, M.; Popowicz, G. M.; Dawidowski, M. 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.* **2025**, *298*, 117979. https://doi.org/10.1016/j.ejmech.2025.117979.

(25) Gumpper, R. H.; Jain, M. K.; Kim, K.; Sun, R.; Sun, N.; Xu, Z.; DiBerto, J. F.; Krumm, B. E.; Kapolka, N. J.; Kaniskan, H. Ü.; Nichols, D. E.; Jin, J.; Fay, J. F.; Roth, B. L. The Structural Diversity of Psychedelic Drug Actions Revealed. *Nat. Commun.* **2025**, *16* (1), 2734. https://doi.org/10.1038/s41467-025-57956-7.

(26) Tokarski, J. S.; Newitt, J. A.; Chang, C. Y. J.; Cheng, J. D.; Wittekind, M.; Kiefer, S. E.; Kish, K.; Lee, F. Y. F.; Borzillerri, R.; Lombardo, L. J.; Xie, D.; Zhang, Y.; Klei, H. E. The Structure of Dasatinib (BMS-354825) Bound to Activated ABL Kinase Domain Elucidates Its Inhibitory Activity against Imatinib-Resistant ABL Mutants. *Cancer Res.* **2006**, *66* (11), 5790–5797. https://doi.org/10.1158/0008-5472.CAN-05-4187.

(27) Birkinshaw, R. W.; Gong, J.; Luo, C. S.; Lio, D.; White, C. A.; Anderson, M. A.; Blombery, P.; Lessene, G.; Majewski, I. J.; Thijssen, R.; Roberts, A. W.; Huang, D. C. S.; Colman, P. M.; Czabotar, P. E. Structures of BCL-2 in Complex with Venetoclax Reveal the Molecular Basis of Resistance Mutations. *Nat. Commun.* **2019**, *10* (1), 2385. https://doi.org/10.1038/s41467-019-10363-1.

(28) Le Trong, I.; Aubert, D. G. L.; Thomas, N. R.; Stenkamp, R. E. The High-Resolution Structure of (+)- *Epi* -Biotin Bound to Streptavidin. *Acta Crystallogr. D Biol. Crystallogr.* **2006**, *62* (6), 576–581. https://doi.org/10.1107/S0907444906011887.

(29) Cowan-Jacob, S. W.; Fendrich, G.; Floersheimer, A.; Furet, P.; Liebetanz, J.; Rummel, G.; Rheinberger, P.; Centeleghe, M.; Fabbro, D.; Manley, P. W. Structural Biology Contributions to the Discovery of Drugs to Treat Chronic Myelogenous Leukaemia. *Acta Crystallogr. D Biol. Crystallogr.* **2007**, *63* (1), 80–93. https://doi.org/10.1107/S0907444906047287.

(30) Molander, G. A.; Gormisky, P. E. Cross-Coupling of Cyclopropyl- and Cyclobutyltrifluoroborates with Aryl and Heteroaryl Chlorides. *J. Org. Chem.* **2008**, *73* (19), 7481–7485. https://doi.org/10.1021/jo801269m.

625

A universal model for drug-receptor interactions – Supplementary Information

Filipe **Menezes**,[1,2,#,*] Adam **Wahida**,[3,#,*] Tony **Fröhlich**,[1,2,*] Phillip **Grass**,[4] Jan **Zaucha**,[1,2] Agata **Barzowska-Gogola**,[6,7] Anna **Czarna**,[7] Andreas **Hochhaus**,[8] Johannes **Nissen-Meyer**,[4] Marcus **Conrad**,[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.*

[6]*Łukasiewicz Research Network, Kraków Institute of Technology, 30-418 Kraków, Poland.*

[7]*Malopolska Centre of Biotechnology, Jagiellonian University, Krakow, Poland.*

[8]*Klinik für Innere Medizin II, Universitätsklinikum Jena, Jena, Germany.*

*these authors contributed equally: F. M., A. W., T. F.

§These authors co-supervised this study: M. C., R. K., 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)

Supplemental Figures

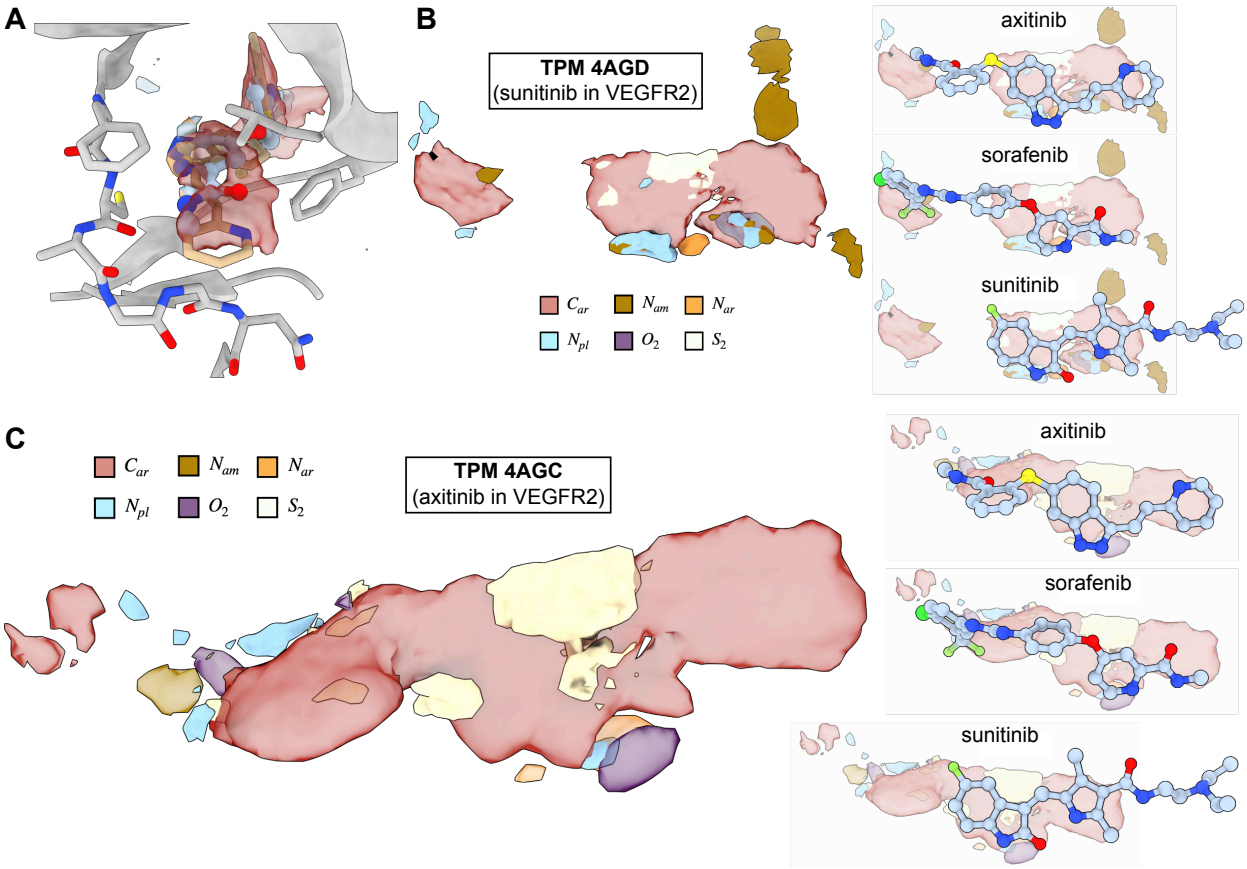


Figure S1. Additional validation of the TPM model for VEGFR crystal structures with sorafenib, sunitinib, and axitinib. A) Overlay of sorafenib and axitinib structures and the TPM predictions for sorafenib in VEGFR. Axitinib's pyridyl ring's nitrogen is not involved in any interaction with the target, 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.

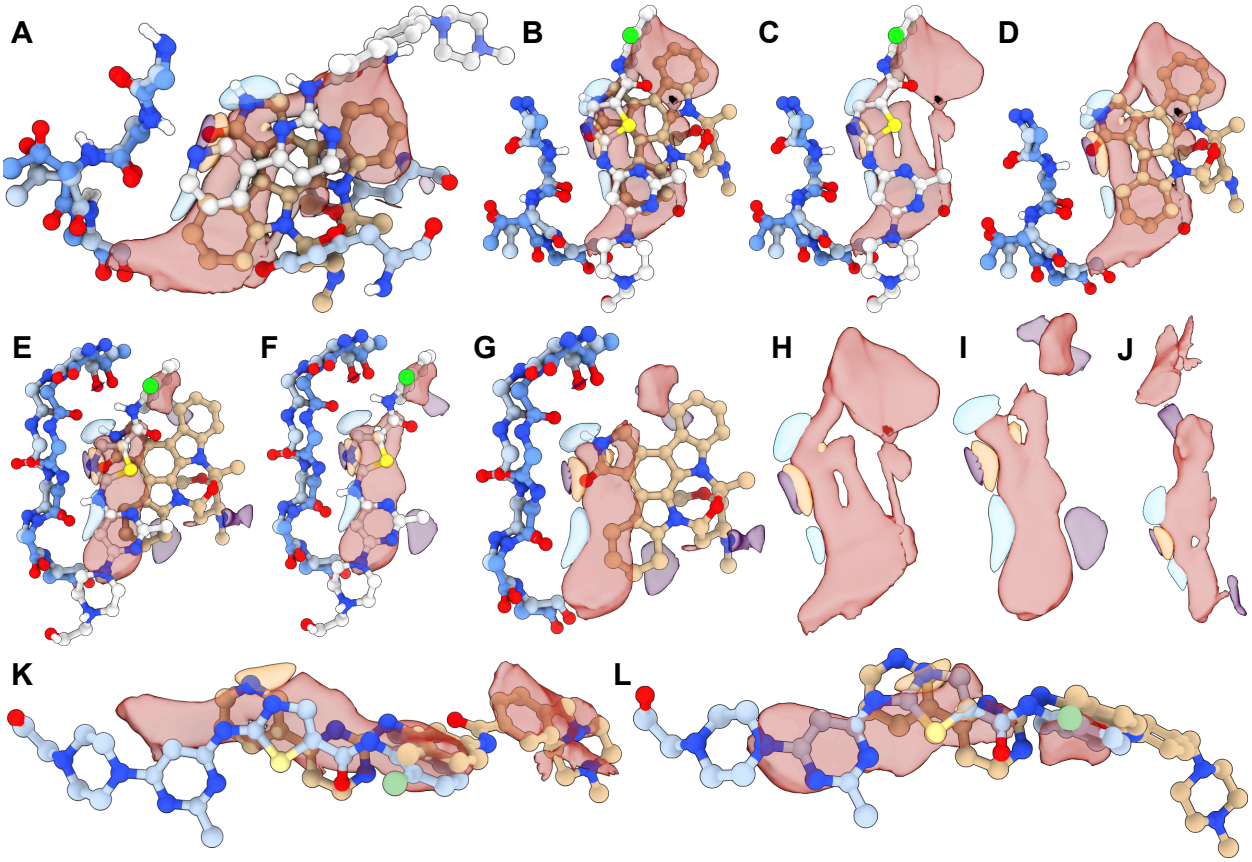


Figure S2: Ligand imprint in experimental crystal structures. A) Overlay of the crystal structure of imatinib in BCR::ABL1 (PDB-ID 2HYY, ligand color code white, protein color code light blue) with that of staurosporine in BCR::ABL1 (PDB-ID 2HZ4, ligand color code brown, protein color code darker blue). 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::ABL1 (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::ABL1. 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::ABL1. Color code identical to that of block B). F) Dasatinib overlaid with the TPMs predicted for 2GQG, the crystal structure of dasatinib in BCR::ABL1. G) Staurosporine overlaid with the TPMs predicted for 2GQG. H) TPMs for 2HZ4 (staurosporine in BCR::ABL1). The latter are to be compared against predictions for I) 2GQG (dasatinib in BCR::ABL1), and J) 2HYY (imatinib in BCR::ABL1). K) Overlay of imatinib (brown) and dasatinib (light blue) with the TPMs for the PDB 2HYY (imatinib in BCR::ABL1). L) Overlay of imatinib (brown) and dasatinib (light blue) with the TPMs for the PDB 2GQG (dasatinib in BCR::ABL1).

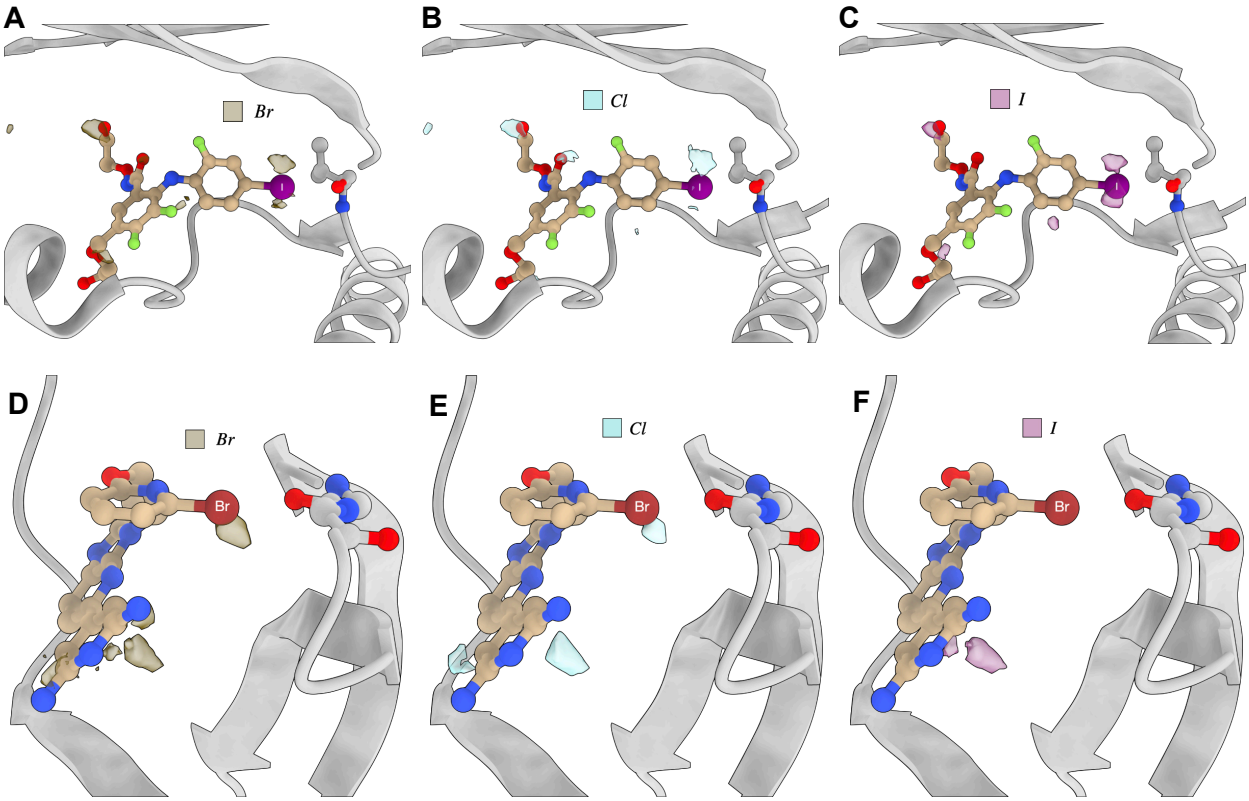


Figure S3. The specificity of halogen signals in TPM predictions. A-C) TPM predictions for the crystal structure 3DV3, the structure of the MEK kinase with halogen-containing inhibitors. A) The bromine, B) chlorine, and C) iodine predictions agree to a good extent, although some signals are specific 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.

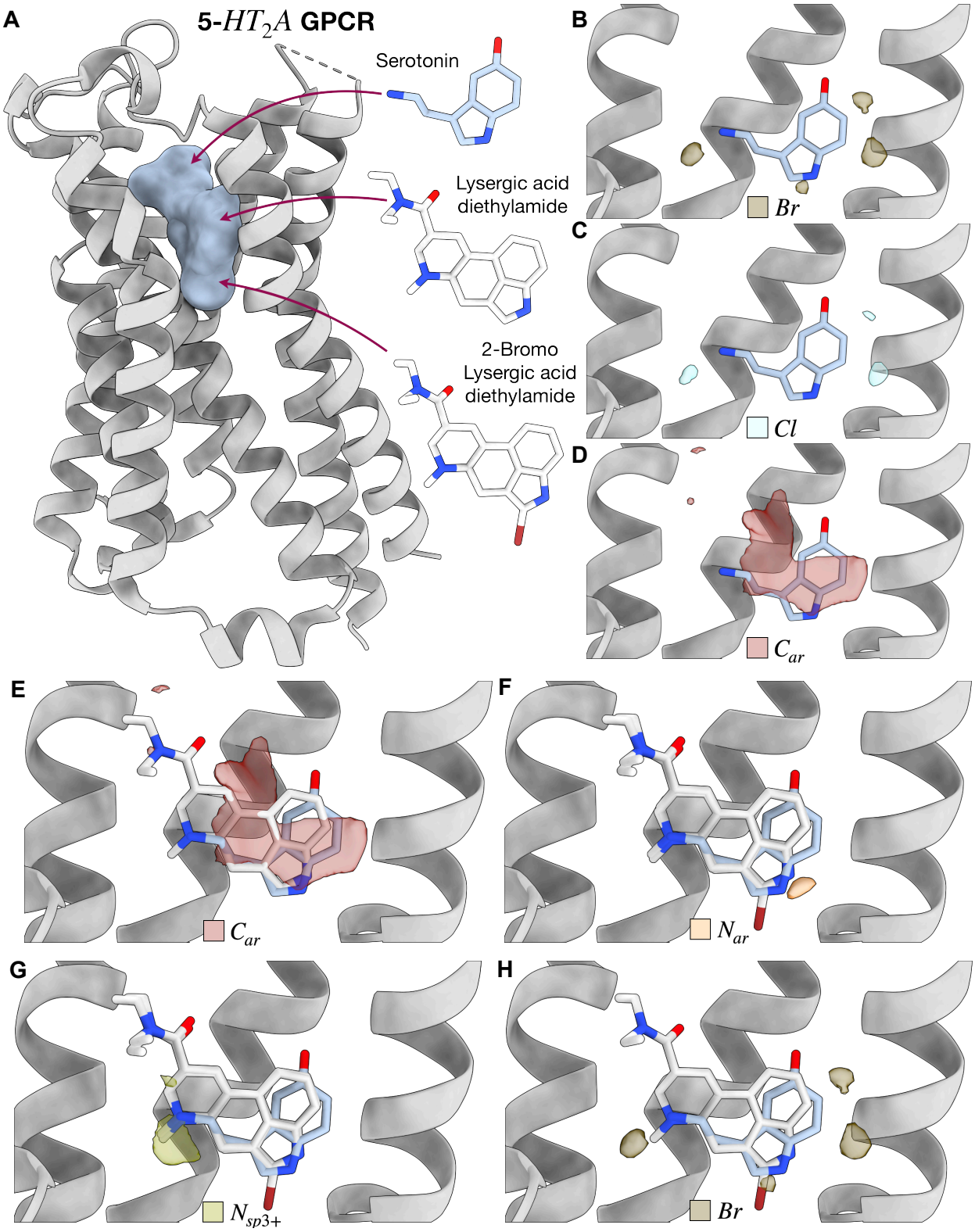
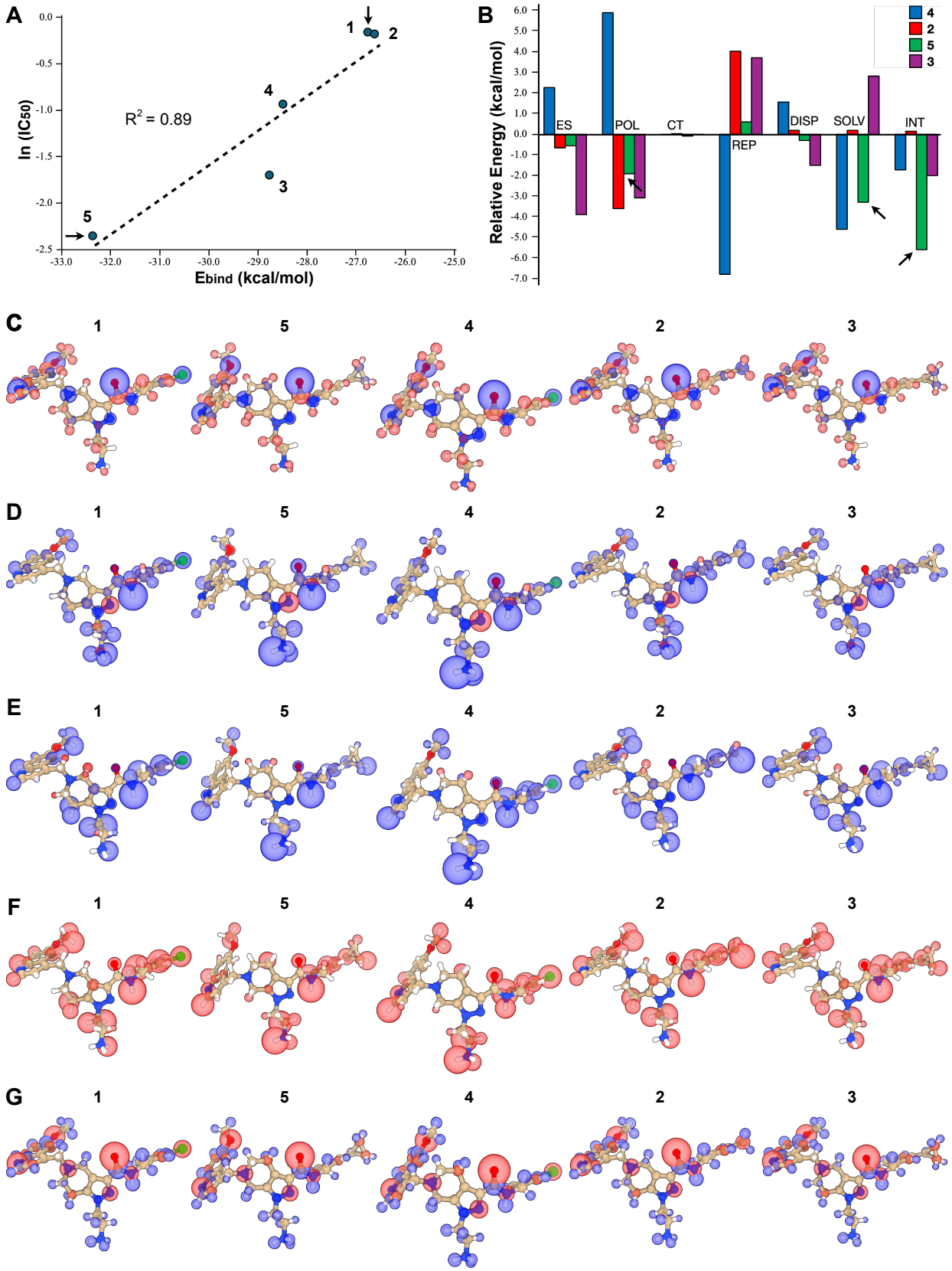


Figure S4. TPM predictions for the 5-HT2A GPCR. A) The receptor's structure with the ligand's density, which might be serotonin, lysergic acid diethylamide, or its 2-bromo derivative. B-D) Overlay of B) bromine, C) chlorine, and D) aromatic carbon densities predicted from the cryoEM structure of the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

677 Similar overlays, but we consider F) aromatic nitrogens or G) ammonium nitrogens instead. H) Overlay of serotonin,
678 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)
R2 = 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

679

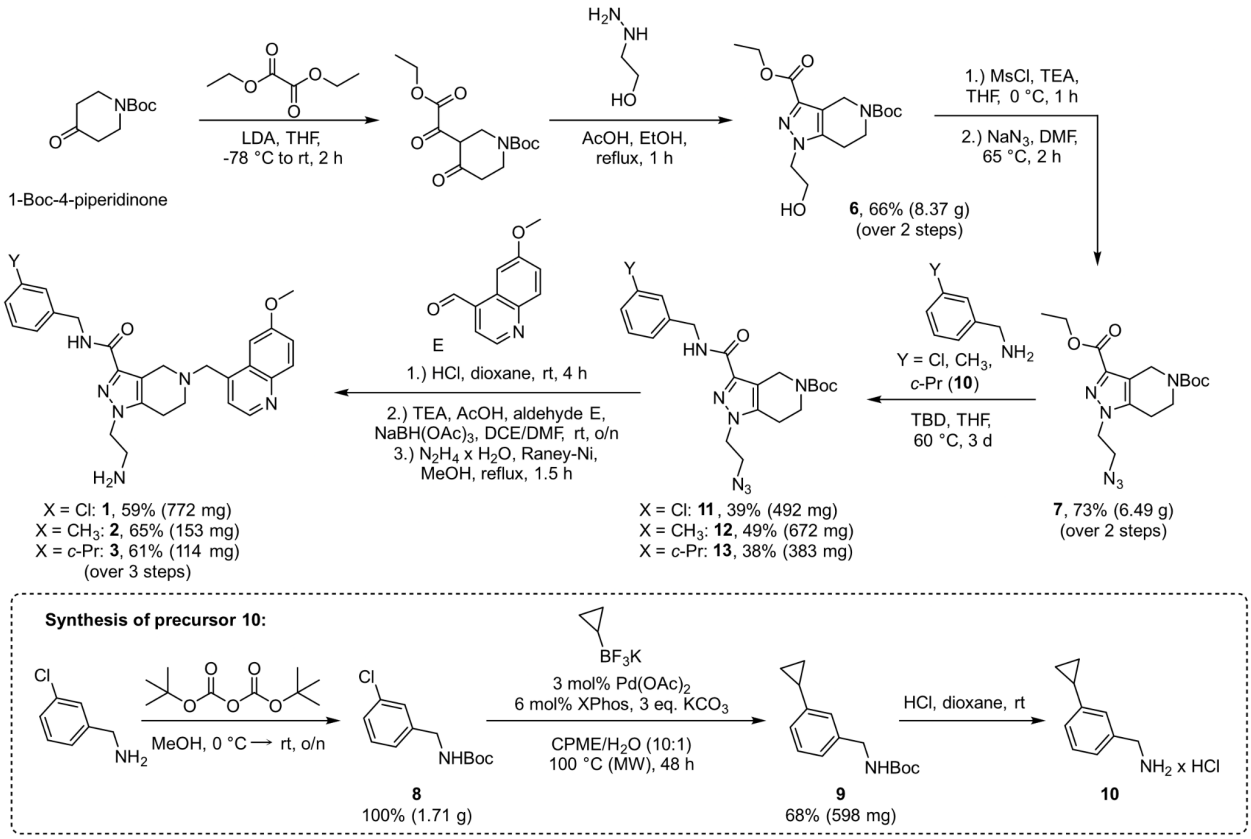
Figure S5. Quantum Mechanical validation of the TPM-generated ligands. A) Correlation plot between the experimentally measured IC50 and the quantum mechanical binding energy. B) The relative Energy Decomposition and Deconvolution Analysis (EDDA) plot, which shows that compound **5**'s improved efficacy 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.

Synthesis of PEX14 inhibitors 1-5

All target compounds were prepared using a slightly modified version of a previously described protocol that was applied to synthesize similar PEX14 inhibitors.[17,18] The optimized synthesis routes are displayed in **Schemes S1** and **S2**, respectively. The detailed synthesis procedures and spectroscopic data are reported in the following paragraphs.

Air and water-sensitive reactions were performed in flame-dried glassware under an argon atmosphere. Solvents used for column chromatography, extractions, and recrystallization were purchased in technical grade and distilled prior to use. Solvents used for reversed phase chromatography and HPLC-MS analyses were purchased from Thermo Fisher Scientific in HPLC-quality. Reagents and dry solvents were purchased from Sigma-Aldrich (Merck), ABCR, Alfa Aesar, Thermo Fisher Scientific, TCI and Carl Roth, and were used without further purification. Analytical thin-layer chromatography (TLC) was performed on silica-coated plates (silica gel 60 F_{254}) purchased from VWR. Compounds were detected by ultraviolet (UV) irradiation at 254 or 366 nm. Manual flash column chromatography was performed using silica gel 60 (particle size: 0.040-0.063 mm) available from VWR. Automated preparative chromatography was performed on a Grace Reveleris Prep purification system using linear gradient elution and Büchi Reveleris Silica 40 µm cartridges for normal-phase and Büchi Reveleris C18 40 µm cartridges for reverse-phase separations. ^{1}H and ^{13}C NMR spectra were recorded at room temperature on a Bruker AV-HD400 spectrometer operating at 400 MHz. NMR peaks are reported as follows: chemical shift (δ) in parts per million (ppm) relative to residual non-deuterated solvent as an internal standard ($CHCl_3$: δ (^{1}H) = 7.26 ppm, δ (^{13}C) = 77.2 ppm; DMSO: δ (^{1}H) = 2.50 ppm, δ (^{13}C) = 39.5 ppm), multiplicity (s = singlet, d = doublet, t = triplet, q = quartet, m = multiplet and br s = broad signal), coupling constant (Hz) and integration. HPLC-UV/MS analyses were performed on a Dionex UltiMate 3000 HPLC system coupled with a Thermo Scientific™ ISQ™ EC Single Quadrupole Mass Spectrometer, using the following method: Thermo Scientific™ Accucore™ RP-MS LC-column (2.1 x 50 mm, 2.6 µm); gradient: 5 to 95% of acetonitrile + 0.1% formic acid v/v in water + 0.1% formic acid v/v over a 5-min period; flow rate: 0.6 mL/min; UV detection at 254 nm. All target compounds exhibited a purity greater than 95%, which was determined by HPLC-UV/MS analyses.

Synthetic procedures and spectroscopic data for PEX14 inhibitors 1-3:



Scheme S1. Synthetic route for the preparation of PEX14 inhibitors **1-3**.

5-(*tert*-butyl) 3-ethyl 1-(2-hydroxyethyl)-1,4,6,7-tetrahydro-5*H*-pyrazolo-[4,3-*c*]pyridine-3,5-dicarboxylate (6). A solution of lithium diisopropylamide (LDA) (2.0 M in THF, 20.6 mL, 41.2 mmol, 1.1 equiv) in dry THF (75.0 mL) was cooled to -78 °C. To this mixture, a solution of 1-Boc-4-piperidone (7.47 g, 37.5 mmol, 1.0 equiv) in dry THF (50.0 mL) was added dropwise under an argon atmosphere. After 30 min of continuous stirring at -78 °C, a solution of diethyl oxalate (6.37 g, 37.5 mmol, 1.0 equiv) in dry THF (50.0 mL) was slowly added. The resulting reaction mixture was then warmed to rt and stirred for an additional 30 min. Subsequently, 2-hydroxyethylhydrazine (2.54 mL, 37.5 mmol, 1.0 equiv), acetic acid (20.0 mL) and EtOH (25.0 mL) were added. After refluxing for 1 h, the reaction was cooled to rt and concentrated in vacuo. The remaining residue was partitioned between EtOAc (100 mL) and a saturated $NaHCO_3$ solution (75.0 mL). The aqueous phase was extracted with EtOAc (3 × 50 mL) and the combined organic extracts were washed with water (50.0 mL) and brine (50.0 mL). After drying over Na_2SO_4, the extract was filtered and concentrated in vacuo. The obtained oily residue was then recrystallized from hexane/CH_2Cl_2 and subsequently dried to yield precursor **6** as an off-white solid (8.37 g, 24.6 mmol, 66%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.59 (s, 2H), 4.37 (q, J = 7.2 Hz, 2H), 4.21 – 4.18 (m, 2H), 4.03 – 4.01 (m, 2H), 3.71 (t, J = 5.8 Hz, 2H), 2.73 (t, J = 5.8 Hz, 2H), 1.47 (s, 9H), 1.37 (t, J = 7.2 Hz, 3H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.1, 154.7, 139.3, 138.2, 117.2, 80.1, 61.2, 60.7, 51.2, 28.2, 21.8, 14.2 ppm. HPLC-MS (ESI): m/z = 362 $[M+Na]^+$; retention time: 3.30 min.

5-(*tert*-butyl) 3-ethyl 1-(2-azidoethyl)-1,4,6,7-tetrahydro-5*H*-pyrazolo[4,3-*c*]pyridine-3,5-dicarboxylate (7). Pyrazole **6** was dissolved in dry THF (50.0 mL) under an argon atmosphere and cooled to 0 °C. After addition of triethylamine (4.97 g, 6.81 mL, 49.1 mmol, 2.0 equiv) and methanesulfonyl chloride (2.66 g. 2.47 mL, 31.9 mmol, 1.3 equiv), the resulting mixture was stirred for 1 h at 0 °C. Afterwards, the reaction mixture was filtered through Celite® and the solvent was removed under reduced pressure. For the second step, the crude mesylate was redissolved in dry DMF (25.0 mL) and sodium azide (4.79 g, 73.7 mmol, 3.0 equiv) was added. The mixture was then heated to 70 °C for 2 h, after which it was quenched by the addition of a saturated $NaHCO_3$ solution (50.0 mL). The aqueous phase was then extracted with EtOAc (3 x 70.0 mL) and the combined organic phases were washed with water (50.0

mL) and brine (50.0 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The crude product was purified via gradient column chromatography (*c*-hexane/EtOAc: 25, 35, 50, 75, 100% EtOac), which yielded the desired product **7** as an off-white solid (6.49 g, 17.6 mmol, 73%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.61 (s, 2H), 4.39 (q, J = 7.2 Hz, 2H), 4.19 (t, J = 5.6 Hz, 2H), 3.78 (t, J = 5.6 Hz, 2H), 3.72 (t, J = 5.8 Hz 2H), 2.73 (t, J = 5.8 Hz, 2H), 1.47 (s, 9H), 1.38 (t, J = 7.15 Hz, 3H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.0, 154.7, 139.3, 139.0, 117.3, 80.1, 60.7, 60.5, 50.8, 48.3, 41.1, 40.1, 28.2, 21.8, 20.8, 14.2 ppm. HPLC-MS (ESI): *m/z* = 365 $[M+H]^+$, 387 $[M+Na]^+$; $^+$; retention time: 3.63 min.

***tert*-Butyl (3-chlorobenzyl)carbamate (8).** To a solution of 3-chlorobenzylamine (1.00 g, 7.06 mmol, 1.0 equiv) in dry MeOH (20.0 mL) Boc anhydride (1.54 g, 7.06 mmol, 1.0 equiv) dissolved in dry MeOH (10.0 mL) was added in a dropwise manner. The reaction mixture was stirred at rt overnight and the next day concentrated in vacuo. The resulting material was pure enough without any further purification and was used directly in the next step (1.71 g, 7.06 mmol, colorless oil, quantitative). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.31 – 7.25 (m, 3H), 7.20 – 7.15 (m, 1H), 4.90 (s, 1H), 4.31 (s, 2H), 1.48 (s, 9H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 156.0, 141.3, 134.6, 130.0, 127.6 (2x C), 125.6, 80.0, 44.3, 28.6 ppm. HPLC-MS (ESI): *m/z* = 242 $[M+H]^+$, retention time: 3.76 min.

Attaching the cyclopropyl group to the benzyl moiety of precursor **8** was accomplished via a Pd-catalyzed cross-coupling reaction that was reported in literature by Molander and Gormisky.[30]

***tert*-Butyl (3-cyclopropylbenzyl)carbamate (9).** A microwave vial was charged with Boc-protected chlorobenzylamine **8** (853 mg, 3.53 mmol, 1.0 equiv), potassium cyclopropyltrifluoroborate (549 mg, 3.71 mmol, 1.1 equiv), and K_2CO_3 (1.46 g, 10.59 mmol, 3.0 equiv). A mixture of CPME/H_2O (10:1) was added as solvent, which was degassed for 30 min via an argon gas stream. XPhos (101 mg, 0.212 mmol, 6 mol%) and $Pd(OAc)_2$ (23.8 mg, 0.106 mmol, 3 mol%) were added to the degassed reaction mixture. Afterwards, the tube was sealed with a cap lined with a disposable Teflon septum and heated at 100 °C for 48 h in a microwave reactor. After this time period, the reaction was cooled to rt, diluted with H_2O (50 mL) and extracted with CH_2Cl_2 (4 x 50 mL). The combined organic layers were washed with brine (50 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The crude product was purified by gradient column chromatography (*c*-hexane/EtOAc: 10, 20% EtOAc) to yield the product as a colorless oil in 68% (598 mg, 2.42 mmol). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.21 (t, J = 7.6 Hz, 1H), 7.05 (d, J = 7.7 Hz, 1H), 7.01 – 6.94 (m, 2H), 4.78 (br s, 1H), 4.27 (s, 2H), 1.88 (tt, J = 8.4, 5.1 Hz, 1H), 1.47 (s, 9H), 0.98 – 0.91 (m, 2H), 0.72 – 0.65 (m, 2H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 156.1, 144.6, 139.0, 128.7, 125.1, 124.8, 124.7, 79.7, 45.1, 28.6, 15.5, 9.4 ppm. HPLC-MS (ESI): *m/z* = 248 $[M+H]^+$, retention time: 3.87 min.

3-cyclopropylbenzylamine hydrochloride (10). Boc-deprotection: Boc-protected 3-cyclopropylbenzylamine **9** (540 mg, 2.18 mmol, 1.0 equiv) was dissolved in dry 1,4-dioxane (15.0 mL) under an argon atmosphere. After the addition of HCl (4.0 M in 1,4-dioxane, 7.64 mL, 30.6 mmol, 14.0 equiv), the resulting reaction mixture was stirred for 4 h at rt. Afterwards, the solvent was removed under reduced pressure and the off-white residual solid was dried in vacuo. The crude product was used without further purification for the follow-up reaction step.

General procedure for the synthesis of precursors 11-13: A solution of pyrazole azide **7** (1.0 equiv), the respective benzylamine (1.2 or 1.0 equiv) and TBD (1.0 or 2.0 equiv) in dry THF was stirred at 60 °C for 4 d. The mixture was concentrated in vacuo and the residue was purified by gradient flash column chromatography (*c*-hexane/EtOAc: 20, 35, 50, 70, 90, 100% EtOAc).

***tert*-Butyl1-(2-azidoethyl)-3-((3-chlorobenzyl)carbamoyl)-1,4,6,7-tetrahydro-5*H*-pyrazolo[4,3-*c*]pyridine-5-carboxylate (11).** Pyrazole azide **7** (1.00 g, 2.75 mmol, 1.0 equiv), 3-chlorobenzylamine (402 µL, 467 mg, 3.30 mmol, 1.2 equiv), TBD (383 mg, 2.75 mmol, 1.0 equiv), dry THF (20.0 mL). Yield: 39% (492 mg, 1.07 mmol), white solid. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.34 (s, 1H), 7.28 – 7.18 (m, 4H), 4.71 (s, 2H), 4.58 (d, J = 6.1 Hz, 2H), 4.13 (t, J = 5.7 Hz, 2H), 3.79 – 3.69 (m, 4H), 2.73 (t, J = 5.6 Hz, 2H), 1.49 (s, 9H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.1, 155.2, 141.1, 140.7, 139.9, 134.6, 130.1, 127.9, 127.7, 126.0, 116.6, 80.4, 61.0, 60.5, 50.7, 48.4, 42.4, 41.4, 40.3, 28.6, 22.1 ppm. HPLC-MS (ESI): *m/z* = 482 $[M+Na]^+$, retention time: 3.87 min.

***tert*-Butyl 1-(2-azidoethyl)-3-((3-methylbenzyl)carbamoyl)-1,4,6,7-tetrahydro-5*H*-pyrazolo[4,3-*c*]pyridine-5-carboxylate (12).** Pyrazole azide **7** (1.13 g, 3.10 mmol, 1.0 equiv), 3-methylbenzylamine (467 µL, 451 mg, 3.72 mmol, 1.2 equiv), TBD (432 mg, 3.10 mmol, 1.0 equiv), dry THF (25.0 mL). Yield: 49% (672 mg, 1.53 mmol), colorless gum. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.23 (t, J = 7.5 Hz, 1H), 7.20 – 7.12 (m, 3H), 7.09 (d, J = 7.5 Hz, 1H), 4.71 (s,

2H), 4.56 (d, J = 5.9 Hz, 2H), 4.11 (t, J = 5.3 Hz, 2H), 3.78 – 3.66 (m, 4H), 2.72 (t, J = 5.8 Hz, 2H), 2.34 (s, 3H), 1.48 (s, 9H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 161.9, 155.3, 141.3, 140.0, 138.6, 138.4, 128.8 (2x C), 128.4, 125.1, 116.7, 80.4, 50.8, 48.4, 43.1, 41.5, 28.6, 27.1, 22.2, 21.5 ppm. HPLC-MS (ESI): m/z = 462 $[M+Na]^+$, retention time: 3.73 min.

***tert*-Butyl1-(2-azidoethyl)-3-((3-cyclopropylbenzyl)carbamoyl)-1,4,6,7-tetrahydro-5*H*-pyrazolo[4,3-*c*]pyridine-5-carboxylate (13).** Pyrazole azide **7** (796 mg, 2.18 mmol, 1.0 equiv), 3-cyclopropylbenzylamine hydrochloride (**10**) (2.18 mmol, 1.0 equiv), TBD (608 mg, 4.37 mmol, 2.0 equiv), dry THF (20.0 mL). Yield: 38% (384 mg, 0.825 mmol), colorless gum. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.23 (t, J = 7.6 Hz, 1H), 7.16 – 7.05 (m, 3H), 6.97 (d, J = 7.7 Hz, 1H), 4.71 (s, 1H), 4.55 (d, J = 5.8 Hz, 2H), 4.10 (t, J = 5.6 Hz, 2H), 3.77 – 3.67 (m, 4H), 2.71 (t, J = 5.7 Hz, 2H), 1.89 (tt, J = 8.4, 5.1 Hz, 1H), 1.48 (s, 9H), 1.00 – 0.90 (m, 2H), 0.73 – 0.66 (m, 2H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.1, 155.3, 144.7, 141.4, 139.9, 138.4, 128.8, 125.7, 125.2, 124.7, 116.6, 80.4, 61.1, 50.8, 48.4, 43.2, 41.5, 28.7, 22.2, 15.5, 9.5 ppm. HPLC-MS (ESI): m/z = 488 $[M+Na]^+$, retention time: 3.86 min.

General procedure for the synthesis of target compounds 1-3:

Boc-deprotection: Pyrazole azide precursor **11**, **12** or **13** (1.0 equiv) was dissolved in dry 1,4-dioxane under an argon atmosphere. After the addition of HCl (4.0 M in 1,4-dioxane, 14.0 equiv), the resulting reaction mixture was stirred for 4 h at rt. Afterwards, the solvent was removed under reduced pressure and the off-white residual solid was dried in vacuo. Reductive amination: The deprotected crude material was suspended in a mixture of dry DCE and DMF under argon and TEA (1.0 equiv) was added. After stirring the reaction for 30 min at rt, 6-methoxyquinoline-4-carbaldehyde (1.0 or 1.1 equiv), AcOH (1.0 equiv) as well as $NaBH(OAc)_3$ (2.0 equiv) were added and the resulting mixture was stirred overnight. The following day, the reaction was quenched with a saturated aqueous solution of $NaHCO_3$ (25 mL) and the aqueous phase was extracted with EtOAc (3 x 35 mL). Afterwards, the combined organic layers were washed with brine (25 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The obtained crude tertiary amine was used for the next step without further purification. Azide reduction: An aqueous slurry suspension of Raney-nickel (1 x tip of a spatula) was washed under argon with MeOH (3 x 5.0 mL). Afterwards, a solution of the crude azide in MeOH was added under argon. After the addition of 50% hydrazine monohydrate solution (3.2 equiv), the reaction mixture was heated until reflux and stirred for 1 h. Afterwards, Raney-nickel was filtered off via Celite®, washed with MeOH (25 mL) and the solvent was removed under reduced pressure. The residue was purified by gradient flash column chromatography (CH_2Cl_2/NH_3 (7.0 M in MeOH): 1.5, 2.5, 3.5, 4, 5, 7% NH_3 (7.0 M in MeOH)) and thereby the desired target compound **1**, **2** or **3** was obtained, respectively.

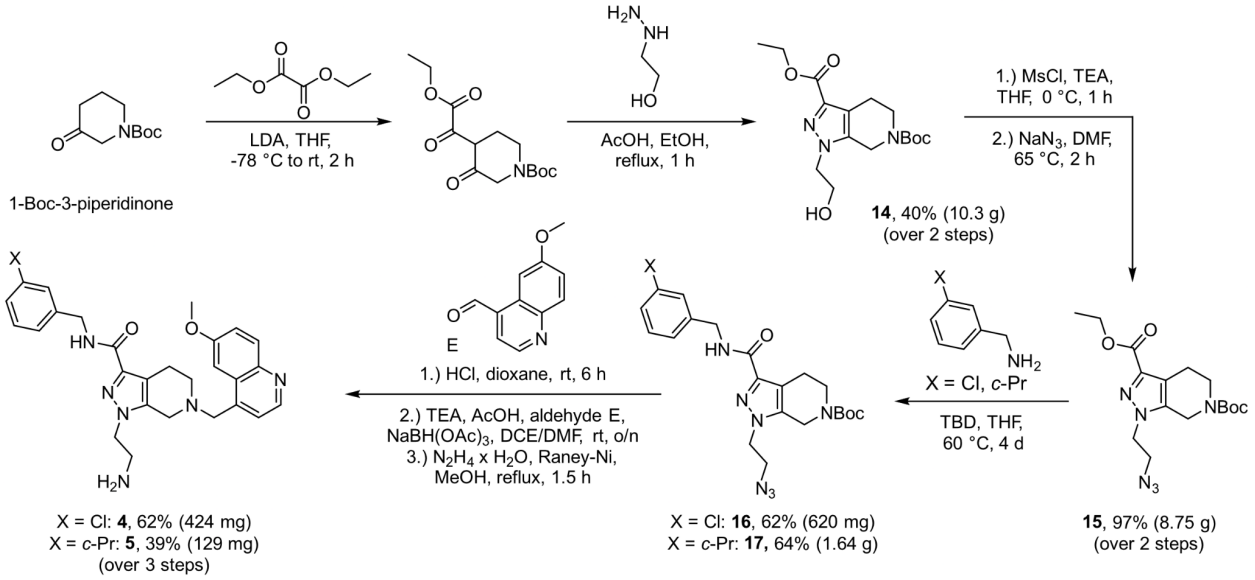
1-(2-Aminoethyl)-*N*-(3-chlorobenzyl)-5-((6-methoxyquinolin-4-yl)methyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[4,3-*c*]pyridine-3-carboxamide (1). Boc-deprotection: Pyrazole azide **11** (1.20 g, 2.61 mmol, 1.0 equiv), HCl (4.0 M in 1,4-dioxane, 9.13 mL, 36.5 mmol, 14.0 equiv), dry 1,4-dioxane (12.0 mL). Reductive amination: TEA (362 µL, 2.61 mmol, 1.0 equiv), 6-methoxyquinoline-4-carbaldehyde (488 mg, 2.61 mmol, 1.0 equiv), AcOH (149 µL, 2.61 mmol, 1.0 equiv), $NaBH(OAc)_3$ (1.11 g, 5.22 mmol, 2.0 equiv), dry DCE/DMF (17.0/3.00 mL). Azide reduction: Hydrazine monohydrate solution (50% in H_2O, 809 µL, 8.35 mmol, 3.2 equiv), Raney-Ni (1 x tip of a spatula), MeOH (20.0 mL). Yield: 59% (772 mg, 1.53 mmol), white foam. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 8.72 (d, J = 4.4 Hz, 1H), 8.02 (d, J = 9.2 Hz, 1H), 7.49 (d, J = 2.8 Hz, 1H), 7.42 (d, J = 4.4 Hz, 1H), 7.36 (dd, J = 9.2, 2.8 Hz, 1H), 7.33 (s, 1H), 7.28 – 7.15 (m, 4H), 4.55 (d, J = 6.2 Hz, 2H), 4.12 (s, 2H), 4.02 (t, J = 5.8 Hz, 2H), 3.98 (s, 2H), 3.90 (s, 3H), 3.14 (t, J = 5.8 Hz, 2H), 2.81 (t, J = 5.6 Hz, 2H), 2.71 (t, J = 5.6 Hz, 2H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.8, 157.7, 147.9, 144.7, 142.6, 140.9, 140.7, 139.2, 134.7, 131.6, 130.1, 128.7, 128.0, 127.7, 126.1, 121.6 (2x C), 117.4, 102.5, 58.9, 55.7, 52.3, 50.7, 49.3, 42.4, 41.9, 22.7 ppm. HPLC-MS (ESI): m/z = 505 $[M+H]^+$, retention time: 2.67 min. HRMS (ESI): m/z calcd. for $[C_{27}H_{29}N_6O_2ClNa]^+$: 527.1938; found: 527.1926.

1-(2-Aminoethyl)-5-((8-methoxyquinolin-4-yl)methyl)-*N*-(3-methylbenzyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[4,3-*c*]pyridine-3-carboxamide (2). Boc-deprotection: Pyrazole azide **12** (213 mg, 0.485 mmol, 1.0 equiv), HCl (4.0 M in 1,4-dioxane, 1.70 mL, 6.80 mmol, 14.0 equiv), dry 1,4-dioxane (7.0 mL). Reductive amination: TEA (67.2 µL, 0.485 mmol, 1.0 equiv), 6-methoxyquinoline-4-carbaldehyde (99.9 mg, 0.543 mmol, 1.0 equiv), AcOH (27.7 µL, 0.485 mmol, 1.0 equiv), $NaBH(OAc)_3$ (206 mg, 0.970 mmol, 2.0 equiv), dry DCE/DMF (7.00/1.00 mL). Azide reduction: Hydrazine monohydrate solution (50% in H_2O, 151 µL, 1.55 mmol, 3.2 equiv), Raney-Ni (1 x tip of a spatula), MeOH (9.00 mL). Yield: 65% (153 mg, 0.316 mmol), white foam. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 8.69 (d, J = 4.4 Hz, 1H), 8.39 (t, J = 6.4 Hz, 1H), 7.94 (d, J = 9.2 Hz, 1H), 7.59 (d, J = 2.8 Hz, 1H), 7.49 (d, J = 4.3 Hz, 1H), 7.40 (dd, J = 9.2, 2.8 Hz, 1H), 7.16 (t, J = 7.5 Hz, 1H), 7.08 – 6.98 (m, 3H), 4.30 (d, J = 6.3 Hz, 2H), 4.14 (s,

2H), 3.98 (t, J = 6.3 Hz, 2H), 3.88 (s, 3H), 3.63 (s, 2H), 2.94 – 2.84 (m, 4H), 2.79 (d, J = 5.4 Hz, 2H), 2.25 (s, 3H) ppm. ^{13}C NMR (101 MHz, DMSO-d$_6$): δ = 162.0, 156.9, 147.6, 143.9, 142.7, 140.0, 139.8, 138.9, 137.1, 130.9, 128.0, 127.9, 127.2, 124.4, 121.3, 121.1, 115.6, 103.0, 57.8, 55.3, 51.9, 49.6, 49.4, 48.6, 41.7, 41.6, 21.7, 21.0 ppm. HPLC-MS (ESI): *m/z* = 485 [M+H]$^+$, 507 [M+Na]$^+$, retention time: 2.32 min. HRMS (ESI): *m/z* calcd. for [$C_{28}H_{32}N_6O_2Na$]$^+$: 507.2484; found: 507.2463.

1-(2-Aminoethyl)-*N*-(3-cyclopropylbenzyl)-5-((6-methoxyquinolin-4-yl)methyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[4,3-*c*]pyridine-3-carboxamide (3). Boc-deprotection: Pyrazole azide **13** (170 mg, 0.365 mmol, 1.0 equiv), HCl (4.0 M in 1,4-dioxane, 1.28 mL, 5.11 mmol, 14.0 equiv), dry 1,4-dioxane (7.00 mL). Reductive amination: TEA (50.6 µL, 0.365 mmol, 1.0 equiv), 6-methoxyquinoline-4-carbaldehyde (71.7 mg, 0.383 mmol, 1.1 equiv), AcOH (20.9 µL, 0.365 mmol, 1.0 equiv), NaBH(OAc)$_3$ (155 mg, 0.730 mmol, 2.0 equiv), dry DCE/DMF (6.00/1.00 mL). Azide reduction: Hydrazine monohydrate solution (50% in H_2O, 113 µL, 1.17 mmol, 3.2 equiv), Raney-Ni (1 x tip of a spatula), MeOH (9.00 mL). Yield: 61% (114 mg, 0.224 mmol), white foam. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 8.69 (d, J = 4.3 Hz, 1H), 8.38 (t, J = 6.3 Hz, 1H), 7.94 (d, J = 9.2 Hz, 1H), 7.59 (d, J = 2.7 Hz, 1H), 7.49 (d, J = 4.4 Hz, 1H), 7.40 (dd, J = 9.2, 2.8 Hz, 1H), 7.14 (t, J = 7.6 Hz, 1H), 7.04 – 6.94 (m, 2H), 6.88 (d, J = 7.7 Hz, 1H), 4.29 (d, J = 6.3 Hz, 2H), 4.14 (s, 2H), 3.98 (t, J = 6.2 Hz, 2H), 3.88 (s, 3H), 3.64 (s, 2H), 2.95 – 2.83 (m, 4H), 2.83 – 2.74 (m, 2H), 1.85 (tt, J = 8.4, 5.9 Hz, 1H), 0.95 – 0.81 (m, 2H), 0.65 – 0.53 (m, 2H) ppm. ^{13}C NMR (101 MHz, DMSO-d$_6$): δ = 162.0, 156.9, 147.6, 143.9, 143.4, 142.7, 140.0, 139.8, 138.9, 130.9, 128.1, 124.7, 124.3, 123.4 (2x C), 121.3, 121.1, 115.6, 103.0, 57.7, 55.3, 51.9, 49.6, 49.4, 41.7, 41.7, 21.7, 14.9, 9.2 ppm. HPLC-MS (ESI): *m/z* = 511 [M+H]$^+$, 533 [M+Na]$^+$, retention time: 2.49 min. HRMS (ESI): *m/z* calcd. for [$C_{30}H_{34}N_6O_2Na$]$^+$: 533.2641; found: 507.2657.

Synthetic procedures and spectroscopic data for PEX14 inhibitors 4-5:



Scheme S2. Synthetic route for the preparation of PEX14 inhibitors **4-5**.

6-(*tert*-Butyl) 3-ethyl 1-(2-hydroxyethyl)-1,4,5,7-tetrahydro-6*H*-pyrazolo[3,4-*c*]pyridine-3,6-dicarboxylate (14). A solution of lithium diisopropylamide (LDA) (2.0 M in THF, 41.4 mL, 82.8 mmol, 1.1 equiv) in dry THF (100 mL) was cooled to -78 °C. To this mixture, a solution of 1-Boc-3-piperidone (15.0 g, 75.3 mmol, 1.0 equiv) in dry THF (50.0 mL) was added dropwise under an argon atmosphere. After 30 min of continuous stirring at -78 °C, a solution of diethyl oxalate (10.2 mL, 75.3 mmol, 1.0 equiv) in dry THF (50.0 mL) was slowly added. The resulting reaction mixture was then warmed to rt and stirred for an additional 30 min. Subsequently, 2-hydroxyethylhydrazine (5.10 mL, 75.3 mmol, 1.0 equiv), acetic acid (25.0 mL) and EtOH (40.0 mL) were added. After refluxing for 1 h, the reaction was cooled to rt and concentrated in vacuo. The remaining residue was partitioned between EtOAc (150 mL) and a saturated $NaHCO_3$ solution (100 mL). The aqueous phase was extracted with EtOAc (3 × 100 mL) and the combined organic extracts were washed with water (75.0 mL) and brine (75.0 mL). After drying over Na_2SO_4, the extract was filtered and concentrated in vacuo. The obtained oily residue was then purified via gradient column chromatography

(*c*-hexane/EtOAc: 10, 25, 35, 50, 75, 100% EtOAc to yield precursor **14** as an off-white solid (10.3 g, 30.3 mmol, 40%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.57 (s, 2H), 4.37 (q, *J* = 7.2 Hz, 2H), 4.22 (t, *J* = 5.0 Hz, 2H), 4.04 (t, *J* = 5.1 Hz, 2H), 3.62 (t, *J* = 5.8 Hz, 2H), 3.58 (br s, 1H), 2.83 (t, J = 5.8 Hz, 2H), 1.47 (s, 9H), 1.38 (t, J = 7.1 Hz, 3H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.3, 154.9, 139.4, 138.2, 118.3, 80.6, 61.3, 60.8, 52.0, 40.4, 28.3, 21.9, 14.3 ppm. HPLC-MS (ESI): *m/z* = 362 $[M+Na]^+$; retention time: 3.26 min.

6-(*tert*-Butyl) 3-ethyl 1-(2-azidoethyl)-1,4,6,7-tetrahydro-6*H*-pyrazolo[3,4-*c*]pyridine-3,6-dicarboxylate (15).

Pyrazole **14** (8.40 g, 24.8 mmol, 1.0 eq) was dissolved in dry THF (50.0 mL) under an argon atmosphere and cooled to 0 °C. After the addition of triethylamine (6.86 mL, 49.5 mmol, 2.0 equiv) and methanesulfonyl chloride (2.49 mL, 32.2 mmol, 1.3 equiv), the resulting mixture was stirred for 1 h at 0 °C. Afterwards, the reaction mixture was filtered through Celite® and the solvent was removed under reduced pressure. For the second step, the crude mesylate was redissolved in dry DMF (25.0 mL) and sodium azide (4.83 g, 75.3 mmol, 3.0 equiv) was added. The mixture was then heated to 70 °C for 2 h, after which it was quenched by addition of a saturated $NaHCO_3$ solution (50.0 mL). The aqueous phase was then extracted with EtOAc (3 × 70.0 mL) and the combined organic phases were washed with water (50.0 mL) and brine (50.0 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The crude product was purified via gradient column chromatography (*c*-hexane/EtOAc: 25, 35, 50, 75, 100% EtOAc), which yielded the desired product **15** as an off-white solid (8.75 g, 24.0 mmol, 97%). ^{1}H NMR (400 MHz, $CDCl_3$): δ = 4.56 (s, 2H), 4.41 (q, *J* = 7.1 Hz, 2H), 4.17 (t, J = 5.5 Hz, 2H), 3.79 (t, J = 5.5 Hz, 2H), 3.62 (t, J = 5.7 Hz, 2H), 2.83 (t, J = 5.7 Hz, 2H), 1.48 (s, 9H), 1.38 (t, J = 7.1, 0.7 Hz, 3H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.5, 155.0, 140.5, 138.3, 118.7, 80.8, 61.0, 51.2, 49.3, 40.5, 28.5, 27.1, 22.1, 14.6 ppm. HPLC-MS (ESI): *m/z* = 387 $[M+Na]^+$, retention time: 3.69 min.

General procedure for the synthesis of precursors 16 and 17: A solution of pyrazole azide **15** (1.0 equiv), the respective benzylamine (1.2 equiv) and TBD (1.0 or 2.0 equiv) in dry THF was stirred at 60 °C for 4 d. The mixture was concentrated *in vacuo* and the residue was purified by gradient flash column chromatography (*c*-hexane/EtOAc: 20, 35, 50, 70, 90, 100% EtOAc).

***tert*-Butyl 1-(2-azidoethyl)-3-((3-chlorobenzyl)carbamoyl)-1,4,5,7-tetrahydro-6*H*-pyrazolo[3,4-*c*]pyridine-6-carboxylate (16).** Pyrazole azide **15** (2.00 g, 5.49 mmol, 1.0 equiv), 3-chlorobenzylamine (805 µL, 933 mg, 6.59 mmol, 1.2 equiv), TBD (764 mg, 5.49 mmol, 1.0 equiv), dry THF (25.0 mL). Yield: 62% (1.57 g, 3.41 mmol), white solid. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.34 – 7.32 (m, 1H), 7.27 – 7.20 (m, 4H), 4.59 – 4.50 (m, 4H), 4.09 (t, *J* = 5.2 Hz, 2H), 3.76 (t, *J* = 5.7 Hz, 2H), 3.62 (t, *J* = 5.5 Hz, 2H), 2.92 (t, *J* = 5.6 Hz, 2H), 1.48 (s, 9H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.3, 142.1, 140.7, 138.5, 134.7, 130.2, 130.1, 128.0, 127.7, 126.1, 117.5, 80.8, 50.7, 49.0, 43.3, 42.5, 28.6, 23.3, 21.9 ppm. HPLC-MS (ESI): *m/z* = 482 $[M+Na]^+$, retention time: 3.80 min.

***tert*-Butyl 1-(2-azidoethyl)-3-((3-cyclopropylbenzyl)carbamoyl)-1,4,5,7-tetrahydro-6*H*-pyrazolo[3,4-*c*]pyridine-6-carboxylate (17).** Pyrazole azide **15** (2.00 g, 5.49 mmol, 1.0 equiv), crude 3-cyclopropylbenzylamine hydrochloride (**10**) (6.59 mmol, 1.2 equiv), TBD (1.53 g, 10.98 mmol, 2.0 equiv), dry THF (25.0 mL). Yield: 64% (1.64 g, 3.52 mmol), colorless gum. ^{1}H NMR (400 MHz, $CDCl_3$): δ = 7.24 – 7.20 (t, *J* = 7.6 Hz, 1H), 7.14 – 7.11 (m, 1H), 7.07 – 7.06 (m, 1H), 6.97 – 6.95 (m, 1H), 4.55 – 4.54 (m, 4H), 4.07 (t, *J* = 5.5 Hz, 2H), 3.74 (t, *J* = 5.6 Hz, 2H), 3.62 (t, *J* = 5.4 Hz, 2H), 2.93 (t, *J* = 5.9 Hz, 2H), 1.91 – 1.85 (tt, *J* = 8.4, 5.1 Hz, 1H), 1.49 (s, 9H), 1.01 – 0.88 (m, 2H), 0.69 (dt, *J* = 6.7, 4.6 Hz, 2H) ppm. ^{13}C NMR (101 MHz, $CDCl_3$): δ = 162.2, 155.2, 144.7, 142.4, 138.5, 138.4, 128.8, 125.7, 125.2, 124.7, 117.2, 80.7, 50.7, 48.9, 43.1, 40.5, 28.6, 21.9, 15.5, 9.4 ppm. HPLC-MS (ESI): *m/z* = 359 $[M-Boc]^+$, retention time: 3.26 min.

General procedure for the synthesis of target compounds 4 and 5:

Boc-deprotection: Pyrazole azide precursor **16** or **17** (1.0 equiv) was dissolved in dry 1,4-dioxane under an argon atmosphere. After the addition of HCl (4.0 M in 1,4-dioxane, 14.0 equiv), the resulting reaction mixture was stirred for 6 h at rt. Afterwards, the solvent was removed under reduced pressure and the off-white residual solid was dried in vacuo. Reductive amination: The deprotected crude material was suspended in a mixture of dry DCE and DMF under argon and TEA (1.0 equiv) was added. After stirring the reaction for 30 min at rt, 6-methoxyquinoline-4-carbaldehyde (1.1 equiv), AcOH (1.0 equiv) as well as $NaBH(OAc)_3$ (2.0 equiv) were added and the resulting mixture was stirred overnight. The following day, the reaction was quenched with a saturated aqueous solution of $NaHCO_3$ (25 mL) and the aqueous phase was extracted with EtOAc (3 x 35 mL). Afterwards, the combined organic layers were

washed with brine (25 mL), dried over Na_2SO_4, filtered and concentrated under reduced pressure. The obtained crude tertiary amine was used for the next step without further purification. Azide reduction: An aqueous slurry suspension of Raney-nickel (1 x tip of a spatula) was washed under argon with MeOH (3 x 5.0 mL). Afterwards, a solution of the crude azide in MeOH was added under argon. After the addition of 50% hydrazine monohydrate solution (3.2 equiv), the reaction mixture was heated until reflux and stirred for 1 h. Afterwards, Raney-nickel was filtered off via Celite®, washed with MeOH (25 mL) and the solvent was removed under reduced pressure. The residue was purified by gradient flash column chromatography (CH_2Cl_2/NH_3 (7.0 M in MeOH): 1.5, 2, 2.5, 3, 4, 5% NH_3 (7.0 M in MeOH)) and thereby the desired target compound **4** or **5** was obtained, respectively.

1-(2-Aminoethyl)-*N*-(3-chlorobenzyl)-6-((6-methoxyquinolin-4-yl)methyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[3,4-*c*]pyridine-3-carboxamide (4). Boc-deprotection: Pyrazole azide **16** (620 mg, 1.35 mmol, 1.0 equiv), HCl (4.0 M in 1,4-dioxane, 4.72 mL, 18.9 mmol, 14.0 equiv), dry 1,4-dioxane (10.0 mL). Reductive amination: TEA (187 µL, 1.35 mmol, 1.0 equiv), 6-methoxyquinoline-4-carbaldehyde (265 mg, 1.42 mmol, 1.1 equiv), AcOH (77.1 µL, 1.35 mmol, 1.0 equiv), $NaBH(OAc)_3$ (571 mg, 2.70 mmol, 2.0 equiv), dry DCE/DMF (12.0/2.00 mL). Azide reduction: Hydrazine monohydrate solution (50% in H_2O, 423 µL, 4.31 mmol, 3.2 equiv), Raney-Ni (1 x tip of a spatula), MeOH (15.0 mL). Yield: 62% (424 mg, 0.840 mmol), white foam. ^{1}H NMR (400 MHz, DMSO-d_6): δ = 8.70 (d, J = 4.3 Hz, 1H), 8.56 (t, J = 6.4 Hz, 1H), 7.94 (d, J = 9.2 Hz, 1H), 7.55 (d, J = 2.8 Hz, 1H), 7.50 (d, J = 4.4 Hz, 1H), 7.40 (dd, J = 9.2, 2.8 Hz, 1H), 7.38 – 7.31 (m, 2H), 7.30 – 7.22 (m, 2H), 4.38 (d, J = 6.3 Hz, 2H), 4.15 (s, 2H), 3.98 – 3.83 (m, 5H), 3.71 (s, 2H), 2.87 (t, J = 6.2 Hz, 2H), 2.83 – 2.74 (m, 4H), 2.55 – 2.46 ppm. ^{13}C NMR (101 MHz, DMSO-d_6): δ = 162.3, 156.9, 147.6, 143.9, 142.7, 142.4, 140.9, 138.8, 132.8, 130.9, 130.1, 128.0, 127.1, 126.5, 126.0, 121.2, 115.0, 102.8, 57.3, 55.4, 52.1, 50.1, 48.3, 41.7, 41.2, 21.5 ppm. HPLC-MS (ESI): *m/z* = 505 $[M+H]^+$, retention time: 2.61 min. HRMS (ESI): *m/z* calcd. for $[C_{27}H_{29}N_6O_2ClNa]^+$: 527.1938; found: 527.1967.

1-(2-Aminoethyl)-*N*-(3-cyclopropylbenzyl)-6-((6-methoxyquinolin-4-yl)methyl)-4,5,6,7-tetrahydro-1*H*-pyrazolo[3,4-*c*]pyridine-3-carboxamide (5). Boc-deprotection: Pyrazole azide **17** (305 mg, 0.656 mmol, 1.0 equiv), HCl (4.0 M in 1,4-dioxane, 2.30 mL, 9.18 mmol, 14.0 equiv), dry 1,4-dioxane (6.00 mL). Reductive amination: TEA (90.9 µL, 0.656 mmol, 1.0 equiv), 6-methoxyquinoline-4-carbaldehyde (135 mg, 0.722 mmol, 1.1 equiv), AcOH (37.5 µL, 0.656 mmol, 1.0 equiv), $NaBH(OAc)_3$ (278 mg, 1.31 mmol, 2.0 equiv), dry DCE/DMF (7.00/1.50 mL). Azide reduction: Hydrazine monohydrate solution (50% in H_2O, 206 µL, 2.09 mmol, 3.2 equiv), Raney-Ni (1 x tip of a spatula), MeOH (10.0 mL). Yield: 39% (129 mg, 0.253 mmol), white solid. ^{1}H NMR (400 MHz, DMSO-d_6): δ = 8.70 (d, J = 4.3 Hz, 1H), 8.38 (t, J = 6.4 Hz, 1H), 7.94 (d, J = 9.2 Hz, 1H), 7.55 (d, J = 2.8 Hz, 1H), 7.50 (d, J = 4.4 Hz, 1H), 7.40 (dd, J = 9.2, 2.8 Hz, 1H), 7.17 (t, J = 7.6 Hz, 1H), 7.07 – 7.00 (m, 2H), 6.90 (d, J = 7.7 Hz, 1H), 4.34 (d, J = 6.3 Hz, 2H), 4.15 (s, 2H), 3.93 – 3.85 (m, 5H), 3.71 (s, 2H), 2.86 (t, J = 6.2 Hz, 2H), 2.81 – 2.71 (m, 3H), 1.87 (tt, J = 8.4, 5.1 Hz, 1H), 0.96 – 0.87 (m, 2H), 0.68 – 0.54 (m, 2H) ppm. ^{13}C NMR (101 MHz, DMSO-d_6): δ = 162.1, 156.9, 147.6, 143.9, 143.4, 142.4, 141.1, 139.9, 138.7, 130.9, 128.1, 128.0, 124.7, 124.3, 123.4, 121.2 (2x C), 114.8, 102.8, 57.3, 55.4, 52.1, 50.1, 48.3, 41.7, 41.7, 21.6, 15.0, 9.2 ppm. HPLC-MS (ESI): *m/z* = 511 $[M+H]^+$, retention time: 2.77 min. HRMS (ESI): *m/z* calcd. for $[C_{30}H_{34}N_6O_2Na]^+$: 533.2641; found: 533.2632.

PEX14 protein expression and purification.

N-terminal domain of *T. brucei*, *T. cruzi* and *H. sapiens* PEX14 (residues 19-84) were cloned into pET21 vector. The plasmids were transformed into E. coli BL21(DE3). The overnight culture was inoculated in autoinduction medium supplemented with 100 µg ml^{-1} kanamycin (15915565,). When the cell density (OD600) reached 0.8, the temperature was lowered to 18°C and the cells were grown overnight. Cells were harvested by centrifugation and dissolved in lysis buffer (50 mM Tris pH 8.0, 300 mM NaCl, 10 mM b-mercaptoethanol, 20 mM imidazole, 10 mg/ml DNAseI, 1 mM AEBSF) and lysed by sonication. The lysates, clarified by centrifugation, were passed over a Ni-NTA agarose resin (Qiagen, Germany) pre-equilibrated with buffer A (50 mM Tris pH 8.0, 300 mM NaCl, 10 mM b -mercaptoethanol, 20 mM imidazole) and the protein of interest was eluted with the same buffer containing 250mM imidazole. Concentrated eluates were further purified on Superdex 75 Hiload 16/60 column (GE Healthcare) in phosphate buffer saline (PBS).

AlphaScreen Assay

The assay was performed according to published protocols (28360328, 31860309). 3 nM N-His-PEX14 from each species (*T. brucei*, *T. cruzi* and *H. sapiens*) was mixed with 10 nM biotinylated PEX5-derived peptide (ALSENWAQEFLA) in a PBS buffer supplemented with 5 mg/mL of BSA and 0.01 % (v/v) Tween-20. 5 µg/mL of

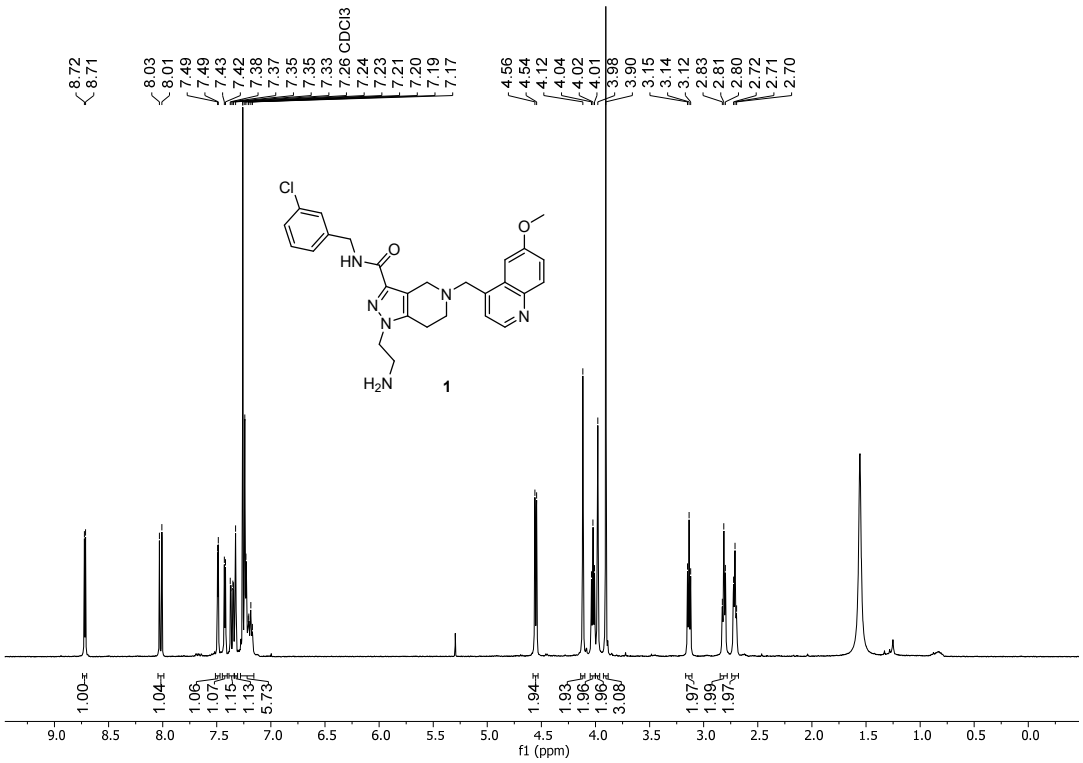
streptavidin donor beads and 5 µg/mL of nickel chelate acceptor beads (PerkinElmer) were added to the mixture. Serial dilutions of the inhibitors (600-0 µM) were prepared in DMSO and added to the protein-peptide mixture keeping constant the concentration of DMSO (5%). Experiments were prepared in triplicates using a liquid handler OT-2 robot (Opentrons). The AlphaScreen assay was performed in white 384-well Optiplates (PerkinElmer, product no. 6007299). The AlphaScreen signals were detected at 520–620 nm using laser excitation at 680 nm, in an EnVision 2102 Multilabel Reader (PerkinElmer). Data were analyzed using Origin Pro 8.5. Experimental points were integrated using Hill sigmoidal fitting fixing the asymptotes at the maximal assay signal (no inhibitor added) and 0, respectively.

Cellular model of Trypanosoma cruzi infection

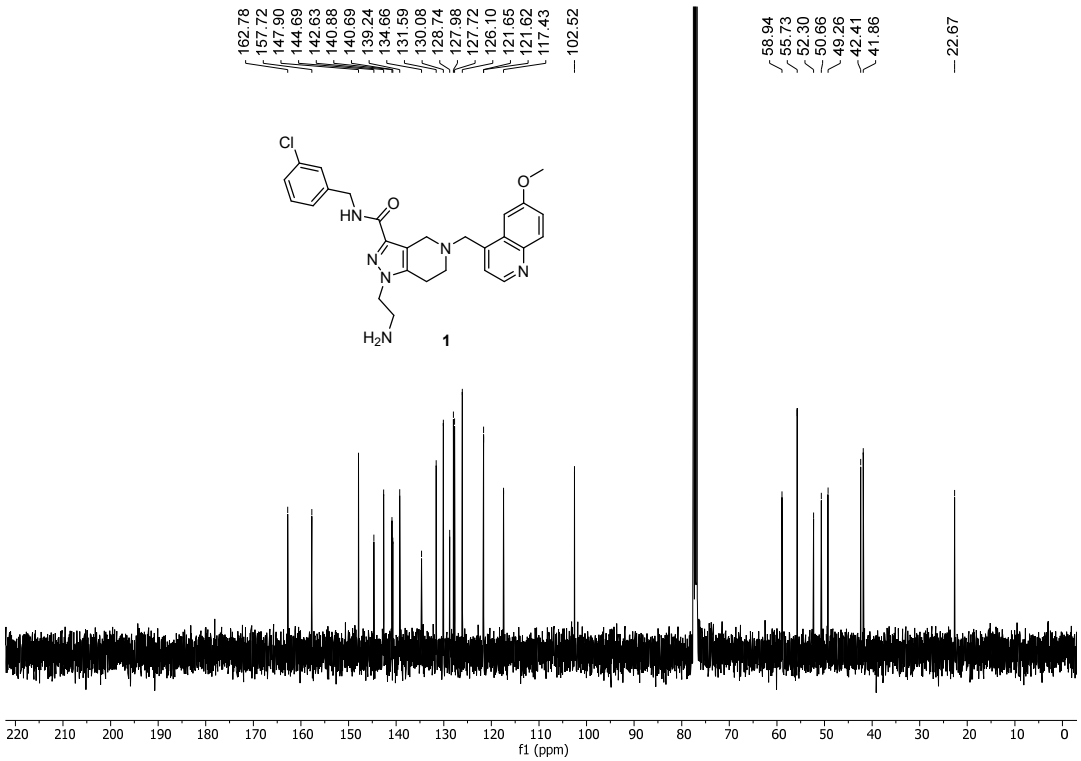
T.cruzi Tulahuen strain C2C4 containing the β-galactosidase (Lac Z) gene[34] as cultivated using Mouse Embryo Fibroblasts (MEF-cells) as host cells, in RPMI 1640 medium supplemented with 1% L-glutamine (200mM) and 10% fetal bovine serum. Cultures were maintained at 37 °C in an atmosphere of 5% CO2. Rat skeletal myoblasts (L-6 cells)[35,36] were cultivated in RPMI 1640 medium supplemented with 1% L-glutamine (200mM) and 10% fetal bovine serum. L-6 Cultures were maintained at 37 °C in an atmosphere of 5% CO2. Assays were performed in 96-well plates containing RPMI 1640 medium supplemented with 1% L-glutamine (200 mM), 10% fetal bovine serum, and $1x10^3$ L-6 cells/well. Plates were incubated at 37 °C under a 5% CO2 atmosphere for 24h. After 24 h, trypomastigote forms of T.cruzi Tulahuen strain C2C4 containing the β-galactosidase (LacZ) gene were added to the wells ($5x10^3$ /well) for infection. Compounds were dissolved in DMSO (10 mg/ml) and 48h after infection serial drug dilutions of eleven 3-fold dilution steps covering 100 to 0.002 µg/mL were prepared in fresh 96 well plates. Benznidazole was used as a reference compound. The medium from the assay plates was removed and replaced by 100 µl of the serial drug dilution plate. After 96h the substrate CPRG/IGEPAL CA 630 (50 µl /well) was added to all wells. A color reaction developed within 2–6h and was read photometrically at 540 nm using the TECAN Sparks (Tecan). IC50 values were calculated from sigmoidal dose response curves by linear regression[37] using Microsoft Excel. Two independent replicates of the assay were performed.

^{1}H and ^{13}C NMR Spectra of Inhibitors 1-5:

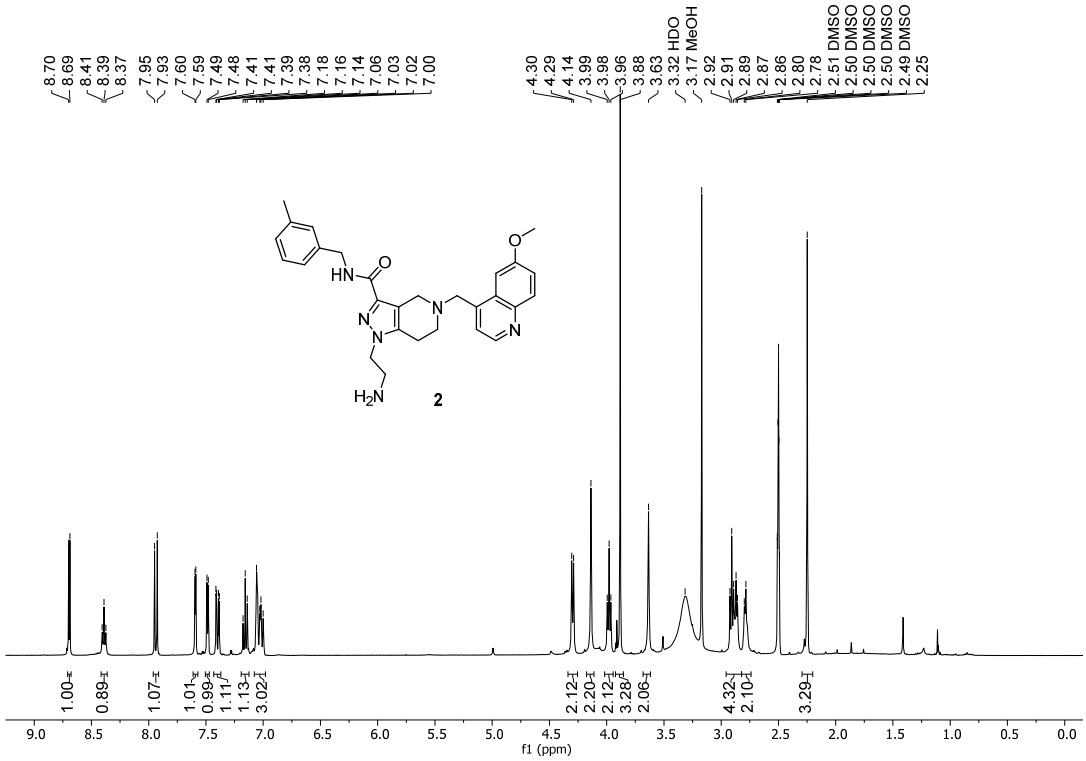
^{1}H NMR spectrum of compound **1** recorded on a Bruker AV-HD400 spectrometer (400 MHz, $CDCl_3$):



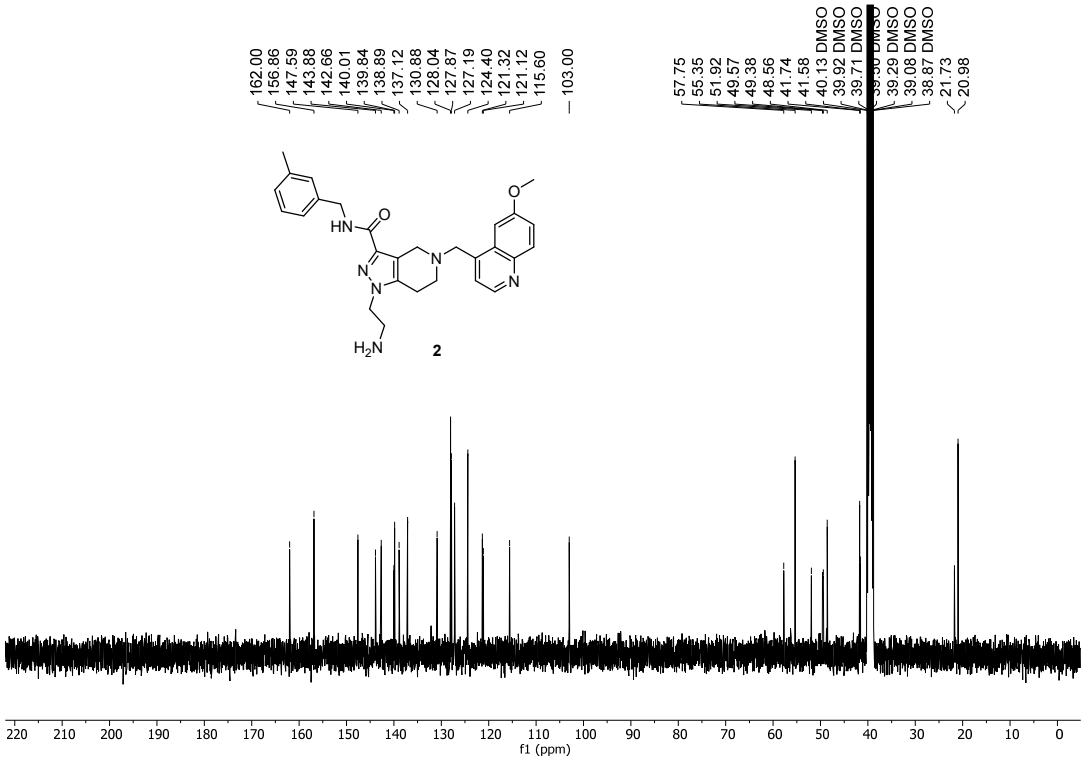
^{13}C NMR spectrum of compound **1** recorded on a Bruker AV-HD400 spectrometer (101 MHz, $CDCl_3$):



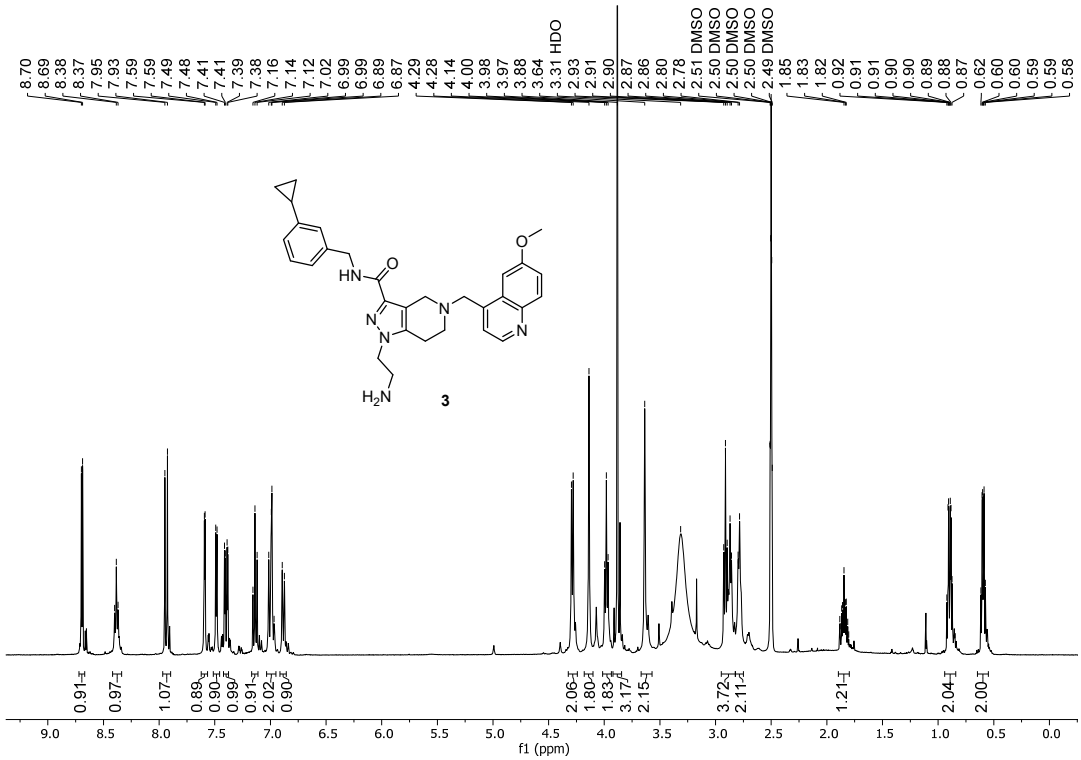
^{1}H NMR spectrum of compound **2** recorded on a Bruker AV-HD400 spectrometer (400 MHz, $CDCl_3$):



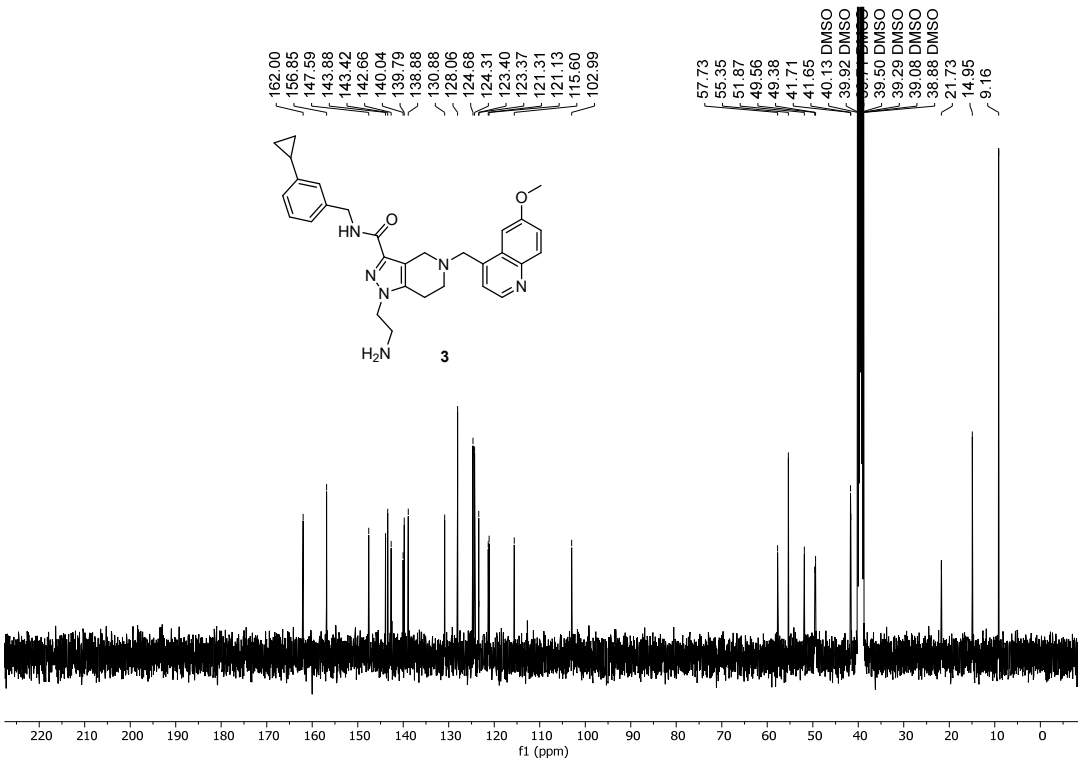
^{13}C NMR spectrum of compound **2** recorded on a Bruker AV-HD400 spectrometer (101 MHz, $CDCl_3$):



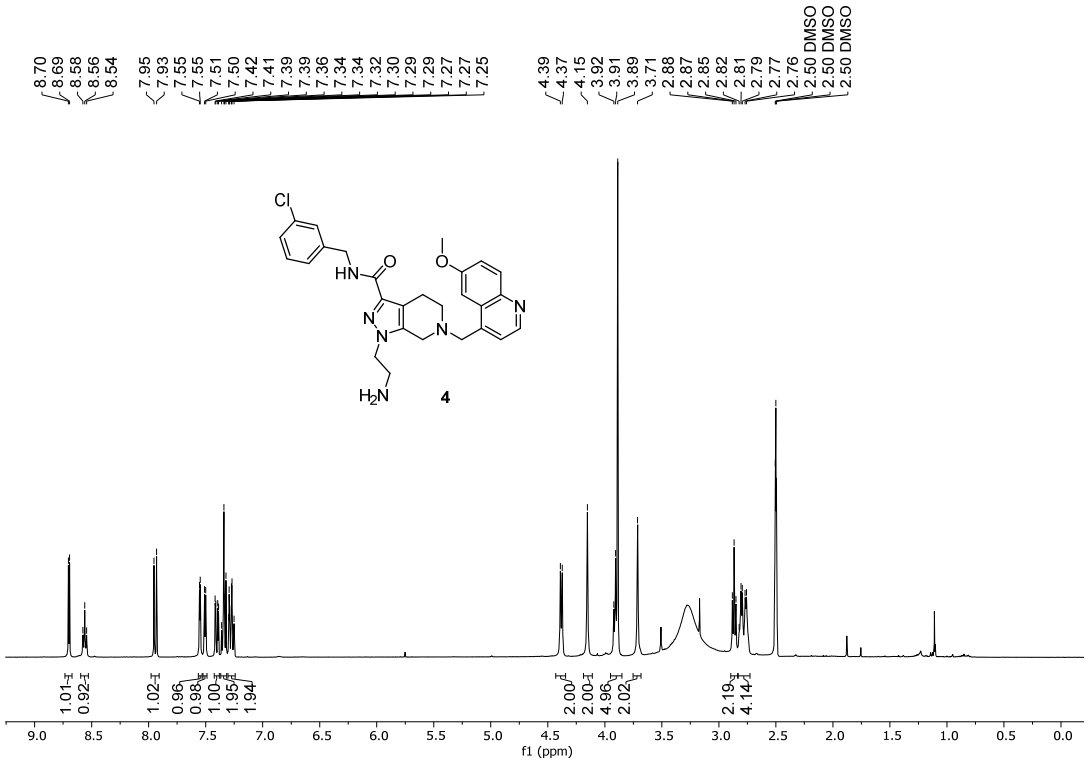
^{1}H NMR spectrum of compound **3** recorded on a Bruker AV-HD400 spectrometer (400 MHz, $CDCl_3$):



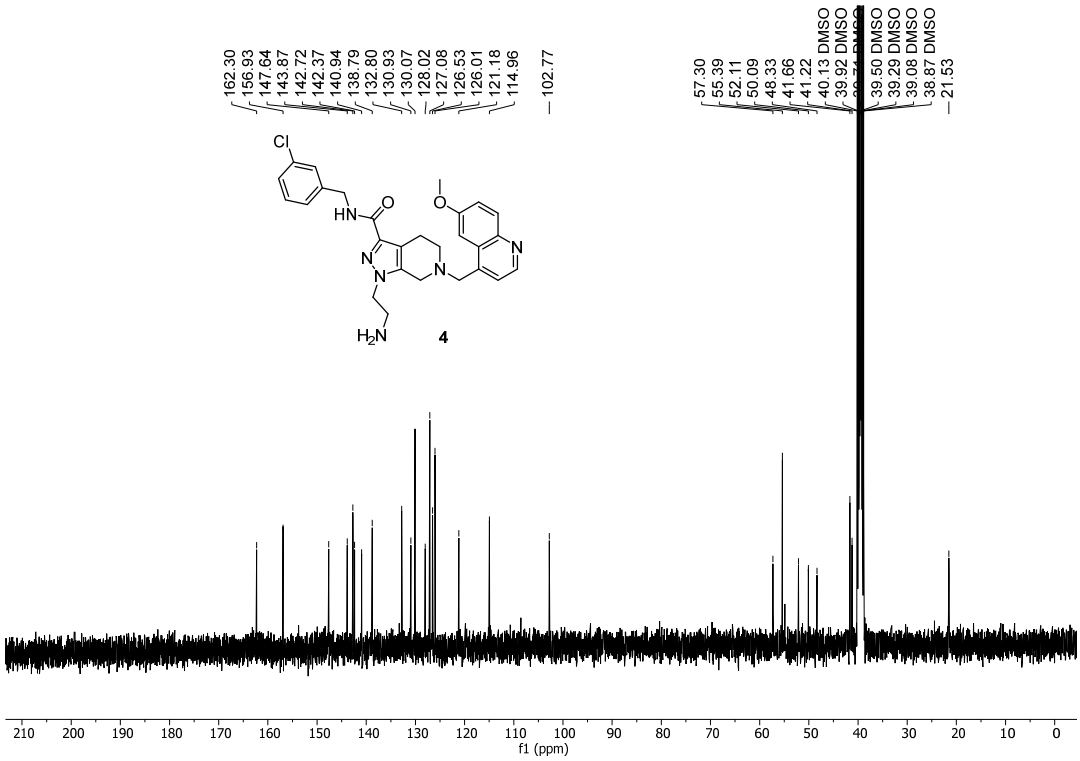
^{13}C NMR spectrum of compound **3** recorded on a Bruker AV-HD400 spectrometer (101 MHz, $CDCl_3$):



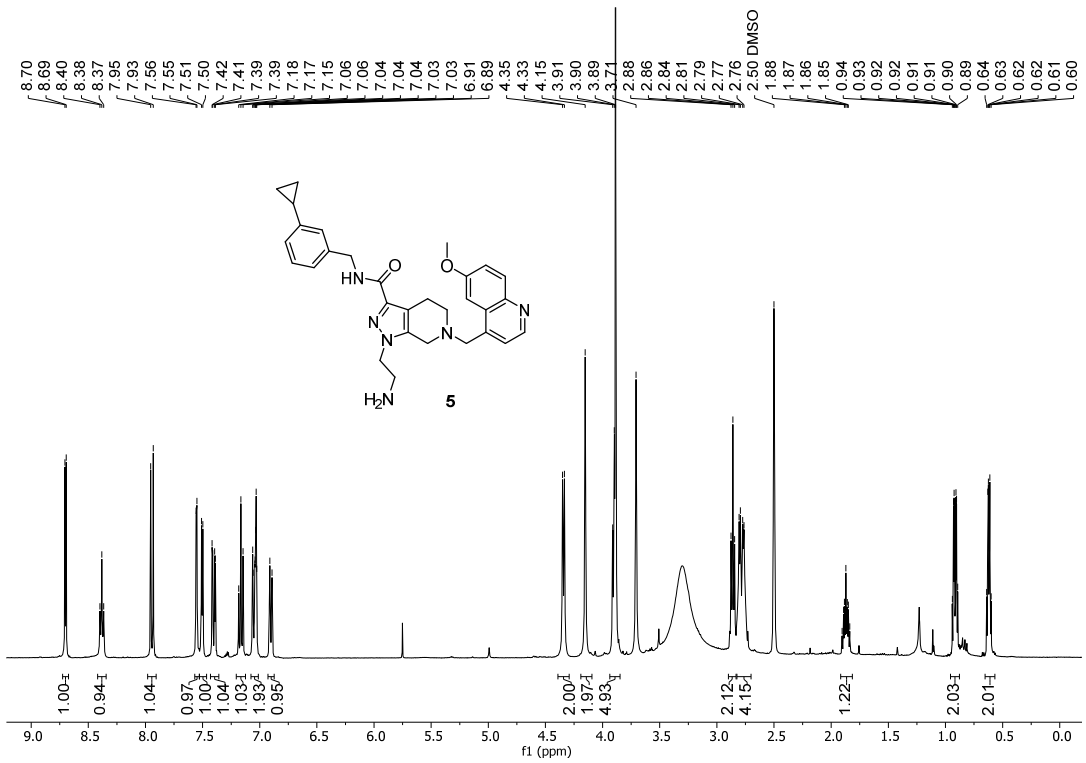
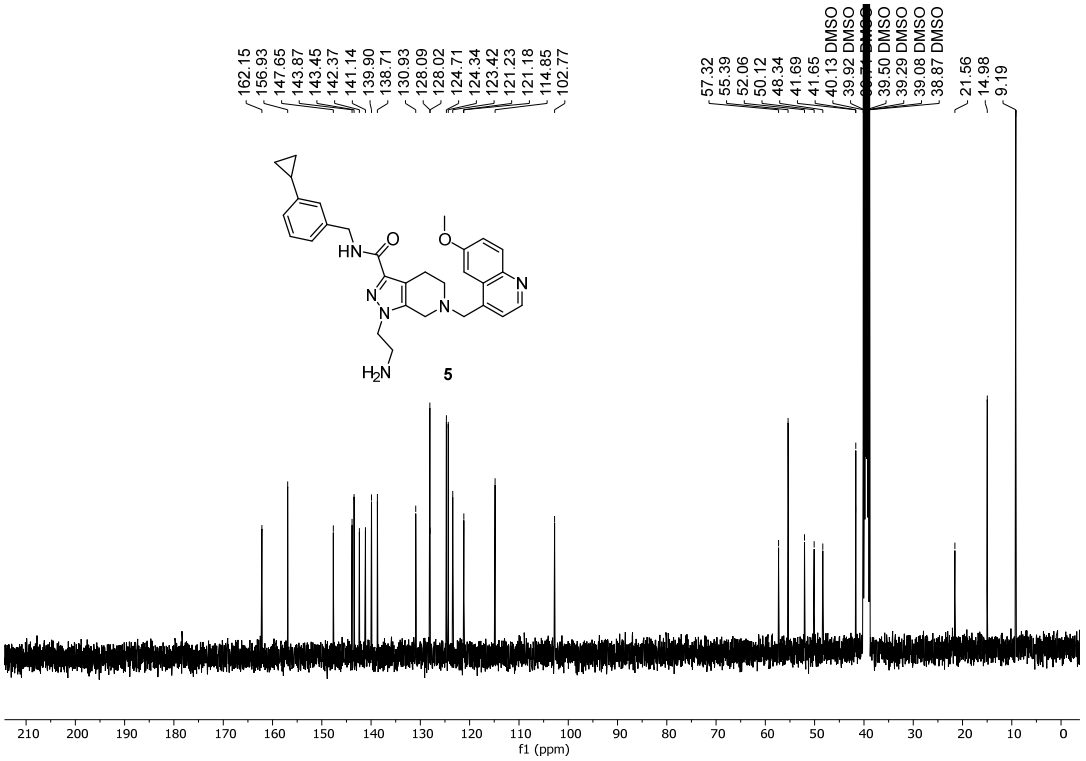
^{1}H NMR spectrum of compound **4** recorded on a Bruker AV-HD400 spectrometer (400 MHz, $CDCl_3$):



^{13}C NMR spectrum of compound **4** recorded on a Bruker AV-HD400 spectrometer (101 MHz, $CDCl_3$):



^{1}H NMR spectrum of compound **5** recorded on a Bruker AV-HD400 spectrometer (400 MHz, $CDCl_3$):



^{13}C NMR spectrum of compound **5** recorded on a Bruker AV-HD400 spectrometer (101 MHz, $CDCl_3$):