nature communications

Article

https://doi.org/10.1038/s41467-024-52146-3

Comprehensive detection and characterization of human druggable pockets through binding site descriptors

Received: 4 March 2024

Accepted: 27 August 2024

Published online: 10 September 2024

Check for updates

Arnau Comajuncosa-Creus[1], Guillem Jorba[1], Xavier Barril[2,3] & Patrick Aloy[1,3]

Druggable pockets are protein regions that have the ability to bind organic small molecules, and their characterization is essential in target-based drug discovery. However, deriving pocket descriptors is challenging and existing strategies are often limited in applicability. We introduce PocketVec, an approach to generate pocket descriptors via inverse virtual screening of lead-like molecules. PocketVec performs comparably to leading methodologies while addressing key limitations. Additionally, we systematically search for druggable pockets in the human proteome, using experimentally determined structures and AlphaFold2 models, identifying over 32,000 binding sites across 20,000 protein domains. We then generate PocketVec descriptors for each site and conduct an extensive similarity search, exploring over 1.2 billion pairwise comparisons. Our results reveal druggable pocket similarities not detected by structure- or sequence-based methods, uncovering clusters of similar pockets in proteins lacking crystallized inhibitors and opening the door to strategies for prioritizing chemical probe development to explore the druggable space.

Ligand binding sites are protein regions that interact with other biochemical entities such as peptides or organic small molecules. The binding process eventually results in a selective modulation of the protein function. Indeed, one of the most successful strategies in conventional drug discovery is to identify, based on the high-resolution three-dimensional structure of binding sites, small molecules that activate or inhibit a protein associated with a disease[1].

Alongside the increasing number of available protein structures in the Protein Data Bank (PDB)[2], structure-based approaches have become a crucial computational framework in early stages of drug development[3,4]. By focusing on ligand binding sites, such strategies enable a rational design and optimization of drugs and reduce the probability of failure of those compounds that reach clinical trials[5]. Protein-small molecule docking is among the most popular structure-based strategies to predict drug-target interactions, and it aims at finding the optimal location and conformation of a given ligand with respect to the receptor binding site[6]. Molecular docking has been successfully applied in proteome-scale studies (e.g. reverse screening[7–9]), but the target-dependent nature of scoring functions prevents the direct comparisons of docking results across different proteins and protein families. Indeed, the design of a universal docking scoring function still remains a challenge[10,11]. Consequently, alternative strategies such as reverse pharmacophore screening, binding site similarity assessment or interaction fingerprint comparison are often employed in proteome-wide analyses[12].

Most of these approaches require the detailed characterization of protein binding sites in a machine-readable format suitable for computational applications, which reasonably allows for the possibility of borrowing featurization techniques from related fields. In fact, characterizing small molecules through numerical vectors encoding

[1]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, Barcelona, Catalonia, Spain. [2]Facultat de Farmàcia and Institut de Biomedicina, Universitat de Barcelona, Barcelona, Catalonia, Spain. [3]Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Catalonia, Spain. e-mail: patrick.aloy@irbbarcelona.org

topological or physicochemical properties is a very common strategy in cheminformatics, and sets the stage for many drug discovery projects founded on the small-molecule similarity principle[13–15]. Likewise, descriptors for larger molecules, such as protein targets, can also be derived, usually gathering features from their amino acid sequences[16]. However, the exploitation of structural data offers a complementary perspective to create protein descriptors and is therefore more promising than the treatment of protein sequences alone. Indeed, biophysical interactions between proteins and ligands occur in very specific areas of protein surfaces (i.e. binding sites) and involve a limited set of residues, which has driven the development of structure-based protein descriptors focused on these particular regions[17].

Pocket descriptors are commonly classified according to the underlying binding site representation they consider, often based on binding site residues (e.g. FuzCav[18], SiteAlign[19]), pocket surfaces (e.g. MaSIF[20]) or explicit interactions with bound ligands or probes (e.g. KRIPO[21], TIFP[22], BioGPS[23]). In addition, and together with the rising interest in deep learning applications in drug discovery[24], data-driven approaches have been designed to derive pocket descriptors borrowing techniques from computer vision (e.g. DeeplyTough[25], BindSiteS-CNN[26]).

Apart from the inherent characterization of binding sites, pocket descriptors provide an excellent means to estimate binding site similarity, which is thus simplified into straightforward vector distance measurements. Binding site comparisons (*aka* pocket matching) have emerged as a promising methodology to move away from the 'one drug-one target-one disease' paradigm[27] by assessing complex studies involving multiple-target drug binding events[28–30]. Binding site similarity is reported to play an important role in the evaluation of ligand promiscuity[31] and in the prediction of protein function, enabling the identification of similar binding sites in proteins having no sequence nor fold similarity[32]. Indeed, the detection of similar binding sites was helpful in several drug repurposing and polypharmacology studies[33–37] and in the prediction of possible distant drug off-targets[38]. In addition, encoding pockets as numerical descriptors entails the possibility of integrating them in a unified framework together with a rich portrait of biochemical entities described in a common vectorial format, such as small molecules, cell lines or diseases[13]. For instance, chemogenomic studies are often addressed by the combination of protein descriptors and molecular fingerprints, usually referred to as proteochemometric (PCM) approaches[39,40].

However, existing methods to generate pocket descriptors exhibit several intrinsic limitations. One of their main drawbacks is the need of co-crystallized ligands to effectively recognize the most relevant biophysical interactions occurring in the binding site, which restricts the applicability domain of such methods to *holo* structures[21,22]. Another important issue related with pocket descriptors is the handcrafted nature of considered binding site representations, often selecting parameters based on specific datasets and performing poorly when used in more general and diverse scenarios[41]. Moreover, several strategies also rely on alignment-dependent comparisons, which makes them particularly useful to provide significant insights into the underlying patterns rationalizing binding site similarity, but also come together with an increased computational cost[19]. In addition, those approaches built upon deep learning algorithms also suffer from lack of interpretability, a well-known problem in the field[42–44]. Finally, the availability of three-dimensional protein structures has traditionally been the main limiting factor in structure-based drug discovery, but this is no longer the case. In the era of accurate protein structure prediction[45–47], where exhaustive collections of predicted structures are available for both relevant organisms[48] and sequences derived from metagenomic studies[49], the structural characterization of proteins is now feasible for essentially any protein sequence of interest. Accordingly, the use of pocket descriptors opens the possibility of characterizing complete proteomes and charting the pocket space in a similar way molecular fingerprints enable the exploration of the chemical space of small molecules[50–52].

To partially overcome the aforementioned limitations of existing pocket descriptors, we exploit the assumption that similar pockets bind similar ligands, which should result in similar rankings in a structure-based virtual screening of small molecules. Indeed, Govindaraj and Brylinski[53] showed that docking scores tended to be more correlated in pockets binding to chemically similar ligands than in pockets binding to dissimilar ligands. This opens the possibility of estimating binding site similarity on the basis of docking rankings and enrichments, as explored by Schmidt and co-workers in their analysis of the human kinome[54]. Moreover, inverse virtual screening (i.e. the screening of a set of targets for a query ligand) has been recently applied to distinguish nucleotide and heme-binding sites from a control set of pockets[55]. In view of these results, we hypothesized that virtual screening could represent a promising strategy to generate pocket descriptors.

Here, we present PocketVec, a strategy to generate interpretable and fixed-length protein binding site descriptors based on the assumption that similar pockets bind similar ligands. Our approach is built upon inverse virtual screening, i.e. the prioritization of a given set of small molecules is expected to be more correlated between similar pockets than between dissimilar ones. We implement and assess the accuracy of our method and the derived pocket descriptors on several predefined benchmark sets. Additionally, we use bound ligands and pocket detection algorithms to comprehensively identify drug binding pockets in experimentally determined and AF2 predicted structures in the human proteome, and derive PocketVec descriptors for all identified pockets. We finally use PocketVec descriptors to exhaustively compare all pockets found in experimental and AF2 structures, explore potential relationships between pocket similarity and small molecule binding and assess a possible complementarity with other sequence and structure-based approaches to demonstrate its potential to find and characterize similar binding sites in unrelated proteins.

Results

It is known that similar proteins tend to bind similar ligands[56], a principle behind many drug discovery projects[12,57,58]. We re-assessed the validity of this principle and found that, indeed, proteins from the same family (e.g. GPCRs) tend to have more similar active compounds than proteins from different families (Fig. S1). However, globally dissimilar proteins showing similar physicochemical and shape properties in their druggable pockets may still bind with similar ligands, which reasonably translates into a more precise and general form of the chemogenomics principle: similar pockets bind similar ligands[59]. PocketVec builds on this observation to generate vector-type descriptors for characterizing protein small molecule binding pockets.

Instead of directly characterizing the shape and physicochemical environment of the protein cavities, we rely on a predefined set of small molecules and assess their potential binding to a given pocket. More specifically, given a three-dimensional protein structure, we first identify possible druggable pockets and we then use computational docking strategies to assess the potential binding of the small molecules. The resulting docking scores are then translated into rankings, which are finally stored in a vector-type format. In this way, each bit of the vector represents the ranking of a predefined molecule, illustrating how good it binds with the pocket of interest compared to all other molecules (Fig. 1a). While the idea is conceptually pretty straightforward, its implementation requires a thorough assessment of the set of used molecules, the docking methodology and the benchmark strategy. The following sections describe our effort to evaluate and optimize each step of the procedure.

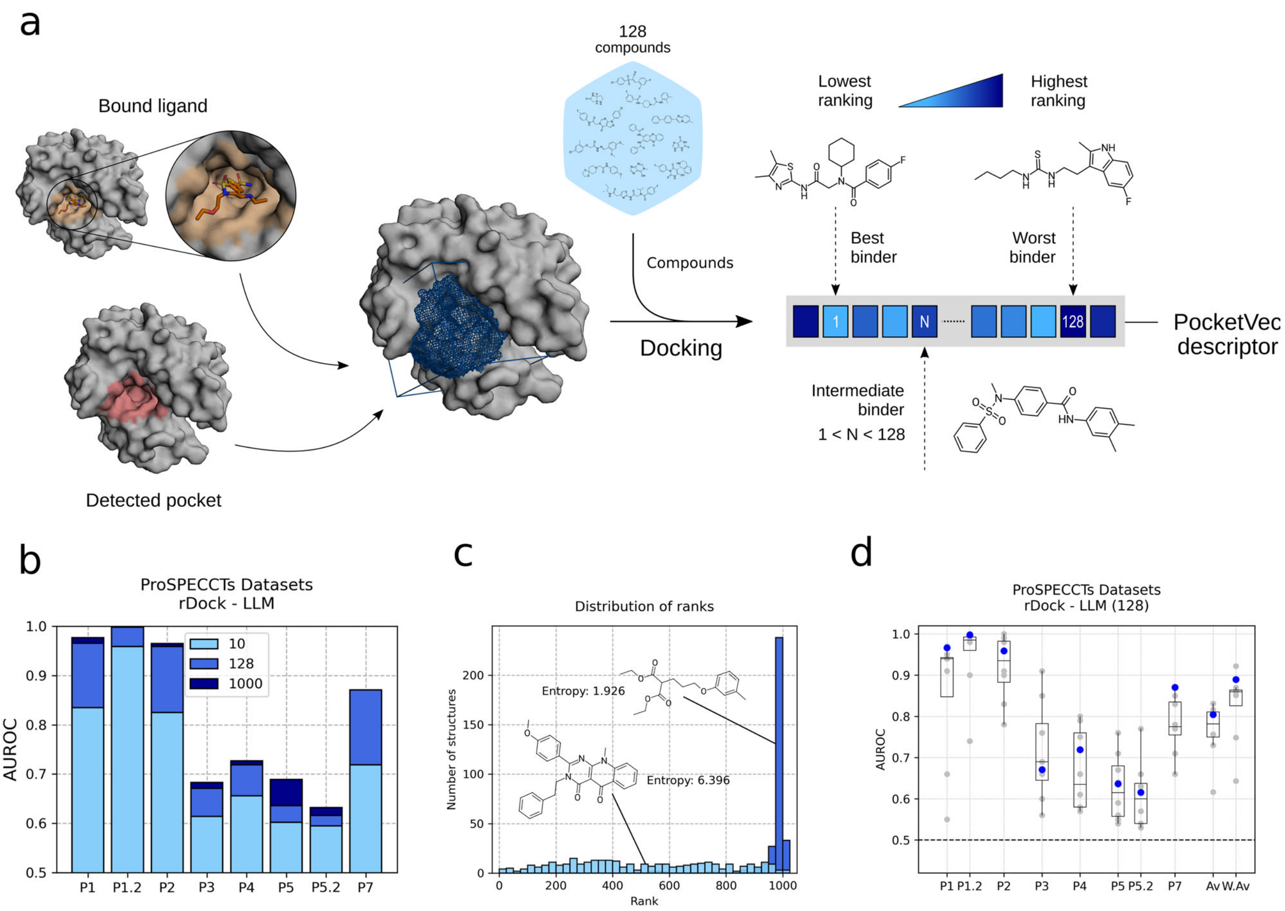


Fig. 1 | PocketVec methodological pipeline and benchmark results. a Given a 3D protein structure, binding site locations are established by the presence of bound ligands or by means of pocket detection algorithms. A predefined set of compounds (128 lead-like molecules in the standard PocketVec pipeline) is docked against the pocket of interest. The corresponding docking scores are then converted into rankings and stored in a vector-type format that serves to characterize the pocket. We refer to those vectors as PocketVec descriptors. **b** Bars indicating the performance (AUROC, y-axis) of our descriptors generated with a varying number of predefined molecules among ProSPECCTs datasets (x-axis, P6 and P6.2 not included. For further details, please see *Methods: Benchmark Set*). Bar color indicates the number of predefined compounds (10, 128 and the complete set of 1000 lead-like molecules, sorted by entropy). These results correspond to the rigid docking (rDock) and LLM combination. All the other combinations with all possible numbers of predefined compounds (from 1 to complete sets) are shown in Fig. S7. **c** Predefined molecules with high and low entropy. The histograms depict the distribution (y-axis) of rankings (x-axis, bin width: 25) for the highest (sky blue) and lowest (dark blue) entropy lead-like molecules in ProSPECCTs P1. Their chemical structures are shown together with the corresponding entropy values. **d** Performances (AUROC, y-axis) of distinct pocket descriptors among ProSPECCTs datasets (x-axis, P6 and P6.2 not included). Gray dots represent the individual performances of existing strategies to derive pocket descriptors ($n = 8$, see Table S1). Box plots indicate median (middle line), 25th, 75th percentile (box), and max and min value within the 1.5 × 25th and 1.5 × 75th percentile range (whiskers). Blue dots indicate the performance of PocketVec descriptors (128 LLM and rDock rigid docking). Av. values represent the average performance among ProSPECCTs datasets for each individual method and W. Av. values weight the average value according to the number of pairs within each dataset. Source data are provided as a Source Data file.

Selection of the methodological pipeline

To find the optimal set of compounds to develop our binding pocket descriptors, we tested two different types of molecules. On the one hand, we used the Glide chemically diverse collection of fragments[60,61], containing 667 compounds with molecular weights in the 50–200 g·mol^{-1} range (Fig. S2). Additionally, we also selected 1000 lead-like molecules (LLM) from the MOE v2019.01 dataset (Chemical Computing Group, Montreal, Canada), exhibiting molecular weights in the 200–450 g·mol^{-1} range (Fig. S2).

We also assessed the performance of two well-established small molecule docking strategies. More specifically, we used rDock[62] and SMINA[63] to run rigid and flexible docking calculations, respectively, under the default parameters.

Finally, to determine the best combination of small molecules and docking methods we relied on ProSPECCTs, a collection of datasets aimed at evaluating the performance of pocket comparison approaches (including pocket descriptors) in a wide range of distinct scenarios[41]. In brief, ProSPECCTs comprises 10 datasets consisting of protein-ligand binding site pairs classified as either similar or dissimilar according to various criteria, including pairs of different structures of the same proteins, proteins harboring artificial mutations in their binding pockets or pairs of unrelated proteins that are able to bind chemically similar ligands (Fig. S3).

Please, see the *Methods* sections *Selection of compound sets, Small molecule docking strategies and Post-docking analysis* for a more detailed description of the methodological pipeline.

PocketVec parameter selection and benchmark

For each ProSPECCTs dataset, we generated PocketVec descriptors for all ligand-defined protein binding sites, and evaluated pocket similarity on the basis of pairwise cosine distances between PocketVec descriptors: the lower the PocketVec distance, the higher the pocket similarity.

First, we obtained results for all possible combinations of docking strategies (rDock - rigid docking and SMINA - flexible docking) and compound collections (1000 lead-like molecules and 667 fragments).

We observed that, in general, the use of LLM and rDock rigid docking provided better results than other combinations (Fig. S4). Positive docking scores (i.e. molecules that could not be accommodated in the binding pocket, see *Methods: Post-docking analysis* for further details) leading to outlier rankings were rare with fragments but more frequent when using LLM (~35% of structures had at least one outlier molecule in the rDock - LLM (1000) combination, Fig. S5). Thus, LLM showed a superior discriminative power with respect to the size of the pocket and, overall, provided higher ranking diversity among all ProSPECCTs pockets (Fig. S6). In fact, we observed how LLM occupied a larger fraction of the binding sites than fragments, while rigid docking might have removed the noise created by similar rankings of the same ligand in different conformations, overall conferring a superior discrimination capacity to the rDock - LLM combination.

Our benchmarks also revealed that some molecules were systematically ranked as very weak binders and were thus not informative for the assessment of pocket similarity. Indeed, we realized that the use of ~100–200 molecules was often sufficient to get competitive performances in most datasets. In order to optimize the set of predefined molecules to derive pocket descriptors, we selected those LLM and fragments presenting high ranking diversity across all ProSPECCTs datasets (see *Methods: Entropy measurements*). In brief, we calculated the Shannon's entropy for each molecule within each dataset and, after intra-dataset normalization, we assigned each molecule an averaged entropy value, which enabled us to prioritize those molecules presenting high ranking diversity (high entropy). Indeed, performances obtained in ProSPECCTs datasets by the 128 most diverse molecules were fairly similar to the original ones using complete sets of compounds (Fig. 1b, Fig. S7). Reducing the number of docked molecules (from 1000/667 to 128) enabled a shorter length of the descriptor, now compatible with other small molecule and biological descriptors derived in the group[64–66], and alleviated the overall computational cost of the methodology. An illustrated example of low- and high-entropy lead-like molecules is shown in Fig. 1c.

Overall, and in view of the benchmark results, we established that the use of rigid docking (rDock) and 128 LLM was the standard and optimal methodology to generate PocketVec descriptors. All results presented along the rest of the manuscript were derived following this strategy. The selected LLM are shown in Fig. S8 and can also be found in our GitLab repository in SMILES and SDF format, together with their commercial names.

PocketVec performance on the ProSPECCTs benchmark sets

The performance of PocketVec descriptors across ProSPECCTs datasets is assessed in terms of the AUROC and is shown in Fig. 1d. We observe that PocketVec descriptors are robust to varying definitions of the same pocket, as determined by different crystallized ligands (P1, AUROC 0.97). When restricting such definitions to chemically similar ligands, the performance is maximal (P1.2, AUROC 1.00). Similarly, our descriptors are robust against protein conformational changes, i.e. protein flexibility (P2, AUROC 0.96), and they are also able to distinguish identical pockets from those altered by 5 artificial mutations leading to changes in physicochemical and shape properties of pocket-lining residues. In these cases, we obtained more modest performances (P3 - physicochemical changes and P4 - both physicochemical and shape changes, AUROCs of 0.67 and 0.72, respectively). Reassuringly though, we observed a significant correlation between the number of artificial mutations (1 to 5) and the corresponding AUROC values (Pearson CC > 0.98, p-value < 0.005 in both P3 and P4, Fig. S9), which confirmed that an increasing number of mutations in the binding site came along with an improved ability to detect such differences using PocketVec descriptors.

We also benchmarked our descriptors in more biologically relevant scenarios where, for instance, pockets binding similar small molecules are found in structurally different proteins. ProSPECCTs includes two datasets to address such cases: P5 includes pocket structures classified into 9 distinct ligand classes (e.g. HEM, ATP, NAD, etc.), and P7 includes a realistic set of binding site pairs reported to be similar in published literature, some of them identified in otherwise unrelated proteins. In these sets, PocketVec descriptors show performances with AUROCs of 0.64 and 0.87, respectively, demonstrating their ability to identify similar pockets in globally dissimilar proteins. It is important to note that two binding sites having identical (or chemically similar) crystallized ligands are not necessarily similar from a PocketVec perspective. Following the logic behind the similarity ensemble approach (SEA[57]), in which targets are quantitatively compared based on the chemical similarity of the ensemble of their ligands, a pair of targets sharing a single active compound may not be significantly similar.

Finally, with the goal of defining a PocketVec distance threshold to classify any pocket pair of interest as either similar or dissimilar, we analyzed the behavior of the Matthew's Correlation Coefficient (MCC) at multiple cut-off values across the different ProSPECCTs datasets (Fig. S10). As expected, each definition of pocket similarity in the ProSPECCTs datasets suggested the choice of a different cut-off distance. For instance, the optimal cut-off in the mutation-related benchmarks (P3 and P4) was around 0.13, while in the most realistic set of binding site pairs reported to be similar in published literature was even lower, around 0.08. However, for general purposes and to minimize false negatives, we selected the distance threshold based on the P1 dataset, where similar pairs were predefined as identical pockets binding chemically distinct ligands whereas dissimilar pairs were unrelated pockets binding different ligands. In this way, we defined 0.17 as our PocketVec threshold distance, which was the value maximizing the MCC in ProSPECCTs P1. However, this threshold should not be regarded as an absolute standard, but rather as a general guideline for classifying a pocket pair of interest as either similar or dissimilar.

For the sake of completeness, we generated results for all combinations of docking methods (rigid - rDock, and flexible - SMINA) and reduced sets of chemical compounds (128 LLM and 128 fragments). The use of rDock and LLM (128) was still the best strategy according to the ProSPECCTs datasets (Fig. S11). Besides, we observed that the selection of compounds leading to the most variable results throughout the different pockets (i.e. high entropy) was, indeed, of great help to distinguish similar from dissimilar pockets in ProSPECCTs P1. This effect was even more pronounced in the fragments, where their lower complexity and molecular weight led to higher promiscuity and redundancy (Fig. S12).

Detailed plots for all ProSPECCTs datasets including ROC Curves, PR Curves and distributions of PocketVec distances, docking scores and pocket volumes are included in our GitLab repository (see *Data availability*).

Comparison with existing strategies

Next, we compared the performance of PocketVec descriptors with state-of-the-art methodologies, as reported in refs. 25,26,41, where the authors benchmarked several pocket comparison tools against the ProSPECCTs datasets, including six strategies based on pocket descriptors (Fig. 1d and Fig. S11). We used the AUROC as the standard performance measure to be able to compare our results with other available methodologies. However, for PocketVec descriptors, we also provide specific values of precision, accuracy, sensitivity, specificity, Matthew´s correlation coefficient and F1 score in each ProSPECCTs dataset (Fig. S13). Overall, PocketVec is the second-best ranked strategy in terms of the weighted average among datasets (0.89) and, indeed, it surpasses the median and the average performance in all ProSPECCTs datasets, apart from P3 (Table S1). The top-scoring method is SiteAlign[19], which is alignment-dependent and based on the projection of residue descriptors into a triangle-discretized sphere that quantifies binding site similarity by minimizing distances between

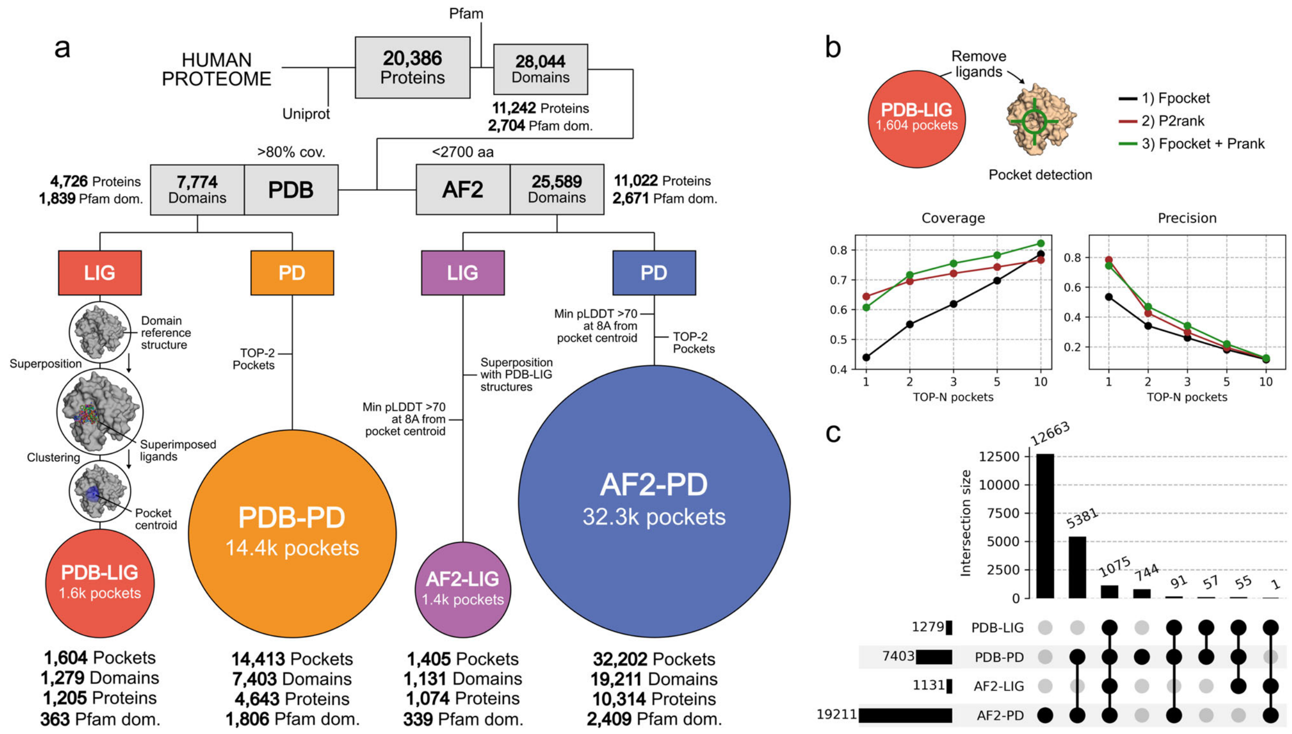


Fig. 2 | Generating PocketVec descriptors for all druggable pockets within human protein domains. a Computational pipeline to gather all available structural data for human druggable pockets included within protein Pfam domains. Starting from the complete human proteome as per Uniprot, we first identified Pfam domains in experimentally determined structures (PDB) and AF2 predicted models. For PDB structures, we identified pockets in *holo* domain structures based on the position of co-crystallized ligands (1604 PDB-LIG pockets, red) and in *apo* domain structures using pocket detection techniques (14,413 PDB-PD pockets, orange). We defined ligand-based pockets on AF2 structures (1405 AF2-LIG pockets, purple) by directly superimposing the location of PDB-LIG pockets, and pockets were also predicted in AF2 models by means of pocket detection algorithms (32,202 AF2-PD pockets, blue). For additional details about the pipeline, please consult the text and the *Methods: Domain-based characterization of the human druggable pockets* section. **b** Benchmark of 3 distinct pocket detection strategies: Fpocket (black), P2rank (brown) and the combination of Fpocket and Prank (green). First, we removed bound ligands from reference PDB structures (PDB-LIG) and we then assessed the performance of the mentioned strategies to detect pockets in ligand-free structures. The plots below represent the evolution of Coverage (proportion of real pockets that are actually detected) and Precision (proportion of detected pockets that are actually real -have a crystallized ligand on PDB-LIG) in terms of the number of best-scoring detected pockets per domain that are considered (1, 2, 3, 5 and 10). **c** UpSet plot representing the intersecting number of domains among pocket sets (1279 domains in PDB-LIG, 7403 domains in PDB-PD, 1131 domains in AF2-LIG and 19,211 domains in AF2-PD). Source data are provided as a Source Data file.

systematically generated cavity fingerprints obtained by moving one binding site with respect to the other. Thus, being specifically developed to compare pockets, it does not provide a unique descriptor for each binding site, which hampers the exploration of the pocket space in the same way molecular fingerprints do for the chemical space of small molecules. Other strategies (e.g[21,22,25,26].) perform similarly to PocketVec throughout ProSPECCTs datasets, but strongly depend on training data and provide pocket embeddings that are difficult to interpret. Thus, overall, PocketVec represents a fast strategy to generate accurate pocket descriptors that overcome the aforementioned limitations, and it shows a higher performance at assessing pocket similarities than most current strategies (Fig. S14).

Comprehensive characterization of druggable pockets in the human proteome

PocketVec descriptors constitute an optimal framework to comprehensively characterize large and diverse sets of small molecule binding sites (apo/holo structures), enabling the navigation across the pocket space of complete proteomes. In view of this, we designed a computational pipeline to generate PocketVec descriptors for all pockets included within human protein domains (Fig. 2a). Please, see *Methods: Domain-based characterization of the human druggable pockets* for a detailed description of the strategy.

In brief, we first retrieved all 20,386 human proteins in UniProt[67]. To avoid working with unstructured or very flexible regions, difficult to model and unlikely to contain druggable pockets, we only kept those protein sequences within autonomous structural units (i.e. domains), as defined in Pfam[68]. Overall, we kept 28,044 domains (2704 unique Pfam domains) in 11,242 human proteins (Fig. S15). The next step was to structurally annotate these domains, for which we used two different strategies. On the one hand, we looked for experimentally determined structures searching the PDB[2], identifying at least one PDB structure for 7774 domains (1839 unique Pfam domains in 4726 proteins). Additionally, we downloaded all human predicted protein structures from AlphaFold DB, obtaining structural models for 25,589 domains (2671 unique Pfam domains in 11,022 proteins). There is an ongoing debate on the use of AF2 models for docking experiments, and whether such models should be refined using i.e. molecular dynamics simulations. However, most studies suggest that the accuracy of AF2 models in docking experiments is comparable to experimentally determined *apo* structures (i.e. without a ligand bound)[69–71]. Indeed, it has been recently shown that the experimental validation rates of small molecule docking is indistinguishable when PDB structures or AF2 models are used[72].

Next, for each domain, we identified the potentially druggable pockets, also using two different strategies. In the first one, that we termed *ligand-based* pocket definition, we searched for small molecules co-crystallized together with the protein domain, and defined druggable pockets as those having bound compounds (HET PDB codes) fulfilling the following criteria: i) being annotated as such in

PDBSUM data, ii) not being one of the 20 naturally occurring amino acids, iii) having more than 6 carbon atoms, to filter out solvent molecules and crystallography-related species and iv) having solvent accessibility ≤0.4 or buriedness ≥15 (Methods: *Domain-based characterization of the human druggable pockets*). In total, we found at least one PDB structure containing small molecules for 1279 domains (363 unique Pfam domains in 1205 proteins). For 503 of these, we only found a single ligand fulfilling our criteria, whereas for 254 of them we could find 10 or more ligands (Fig. S16). Then, to compile the list of unique ligand-defined pockets, we chose a reference PDB structure for each protein domain, and superimposed all domain structures with the corresponding bound ligands onto the reference. We then used a single-linkage clustering strategy to merge into a single pocket all those ligands whose centroids were at a distance ≤5 Å while maintaining the maximum distance between the global centroid of the cluster and the centroids of the individual compounds ≤18 Å. We considered the final global cluster centroids as the pocket centroids. Overall, we found 1604 ligand-defined pockets in 1279 protein domains (363 unique Pfam domains in 1205 proteins). We named this set of pockets PDB-LIG. To apply the same criteria to those structures modeled with AF2, for each domain, we superimposed the reference PDB structure onto the AF2 model, and transferred the location of the identified PDB-LIG pockets accordingly. We only considered those pockets having a pLDDT value > 70 for all residues at a distance ≤8 Å from the pocket centroid. In total, we identified 1405 pockets in 1131 domains (339 unique Pfam domains in 1074 proteins), and named this set *AF2-LIG*. As a complementary strategy, and to increase the overall coverage of the human pocketome, we attempted a de novo identification of druggable pockets. To find the most accurate strategy to predict druggable pockets, we assessed the accuracy of different methods at identifying the PDB-LIG pockets defined above. In brief, we first removed bound ligands from reference *holo* structures and used Fpocket[73] and P2rank[74], to detect pockets in ligand-free domain structures (Methods: *Domain-based characterization of the human druggable pockets*). Our benchmark showed that the best strategy to detect binding sites in ligand-free structures was the combination of Fpocket[73], for pocket detection, and Prank[74] to score them. Using this combination, and considering the top-2 best scored pockets for each domain, we were able to detect 72% of the real pockets while 47% of detected pockets were indeed real (Fig. 2b, coverage and precision, respectively). Thus, for each domain, we ran Fpocket on the PDB reference structure to identify potential pockets, we ranked them by means of Prank, and we kept the top-2 ranked pockets per domain. Overall, this accounted for a total of 14,413 predicted pockets in 7403 domains (1806 unique Pfam domains in 4643 proteins). We named this set *PDB-PD*. We then used the same strategy and criteria as before to detect pockets onto the predicted AF2 domain models, annotating a total of 32,202 pockets in 19,211 domains (2409 unique Pfam domains in 10,314 proteins). We named this set *AF2-PD*.

Robustness in the detection of druggable pockets

Globally, using the different strategies to structurally annotate human protein domains and to identify validated and potential druggable pockets, we compiled 1604 PDB-LIG, 1405 AF2-LIG, 14,413 PDB-PD and 32,202 AF2-PD druggable pockets, in 20,067 domains (Fig. 2a and Fig. 2c). The significant added value of de novo methodologies is very apparent. Indeed, for 18,788 domains (93.6%), all pockets were identified only with pocket detection strategies and, among these, 12,663 domains (63.1%) exclusively featured pockets on AF2 models (Fig. 2c).

Interestingly, when we assessed the methodological robustness of pocket predictions using a subset of 1000 PDB-PD domains, we observed that results slightly differed after translating and rotating structures. This effect was observed for both PDB and AF2 structures in a consistent manner: only ~85% and ~86% of detected pockets were evenly identified after rotation and translation, respectively (top-2 pockets, Fig. S17a). However, in the pockets identified regardless of variations in the initial structures, the scoring was very consistent both in PDB structures and AF2 models (Pearson CC ~0.98). We also evaluated the coherence between pockets detected in PDB and AF2 structures, finding that only ~59% and ~49% of detected pockets were evenly identified in PDB and AF2 structures with respect to a reference PDB structure. However, as when comparing discrepancies due to different initial orientations, in the pockets identified both in PDB and AF2 structures the scoring was quite robust, with Pearson CC of ~0.88 and ~0.62 for PDB and AF2 models, respectively (Fig. S17b).

Additionally, we also explored potential differences in the physicochemical properties between the real (ligand-defined) and predicted (detected) pockets, comparing their volumes and buriedness values (see *Methods: Domain-based characterization of the human druggable pockets*). In this case, as the only remarkable difference, we found that real pockets identified from bound ligands tended to be smaller than those predicted in the PD framework, with average volumes of ~3000 Å^3 and ~3800 Å^3 for LIG and PD pockets, respectively (Fig. 3a). As expected, we reached the same conclusion with buriedness values (Fig. S18), since pocket volume and buriedness were indeed negatively correlated (Fig. S19).

Altogether, these results underscore some limitations in the pocket detection strategy, revealing slight variations due to the orientation of the initial structures, a stronger dependence on structural variability and the production of predicted pockets whose physicochemical properties (e.g. volume) diverge from known pockets.

Systematic generation of PocketVec descriptors

Once we had identified human druggable pockets with four different approaches, we systematically generated PocketVec descriptors for all of them, using the strategy described above (Fig. 1a). The high number of pockets explored enabled an exhaustive analysis of potential dependencies between the lead-like molecules used to build the descriptors and the characterization of druggable pockets. Reassuringly, we did not find any correlation between docking rankings and molecular properties of docked compounds (e.g. molecular weight or number of heavy atoms, Fig. S20 and Fig. S21, respectively). On the contrary, we observed that most molecules exhibited a complete range of rankings (from 1 to 128), although the ranking distributions showed that some molecules tended to bind with good scores in many pockets while others were mostly downranked (Fig. S22). Indeed, this tendency was observed in all pocket definition strategies, and we found significant correlations between the average docking rankings for docked lead-like molecules among different pocket sets (Fig. 3b). Such correlations showed a perfect agreement between PocketVec descriptors generated through PDB and AF2 structures, but exposed slight differences between LIG and PD results (Fig. 3b, Pearson's CC < 0.88). In line with this, we found that 88.0% and 86.8% of the PocketVec descriptors generated for PDB-LIG and AF2-LIG, respectively, showed no lead-like molecules with positive docking scores (i.e. outlier molecules not fitting in the pocket, see *Methods: Post-docking analysis*), while the fraction was considerably higher in PDB-PD and AF2-PD sets, with 97.0% and 98.1% of the descriptors, respectively (Fig. 3c). This result was consistent with the differences observed in pocket volumes (Fig. 3a). Reassuringly, we found that smaller pockets (<3000 Å^3) tended to be more prone to exhibit outlier molecules than bigger pockets (Fisher exact test, OR > 70 and p-value < 10^{-45} for all pocket sets).

To assess if outlier compounds could have a significant effect in the generation of poor PocketVec descriptors, we randomly inserted outlier values in the descriptors and computed the fraction of originally dissimilar pocket pairs (PocketVec distance >0.17) that were incorrectly labeled as similar (distance <0.17) due to an increasing number of outlier values. We observed that the insertion of up to 80 outlier molecules (out of 128) did not significantly compromise

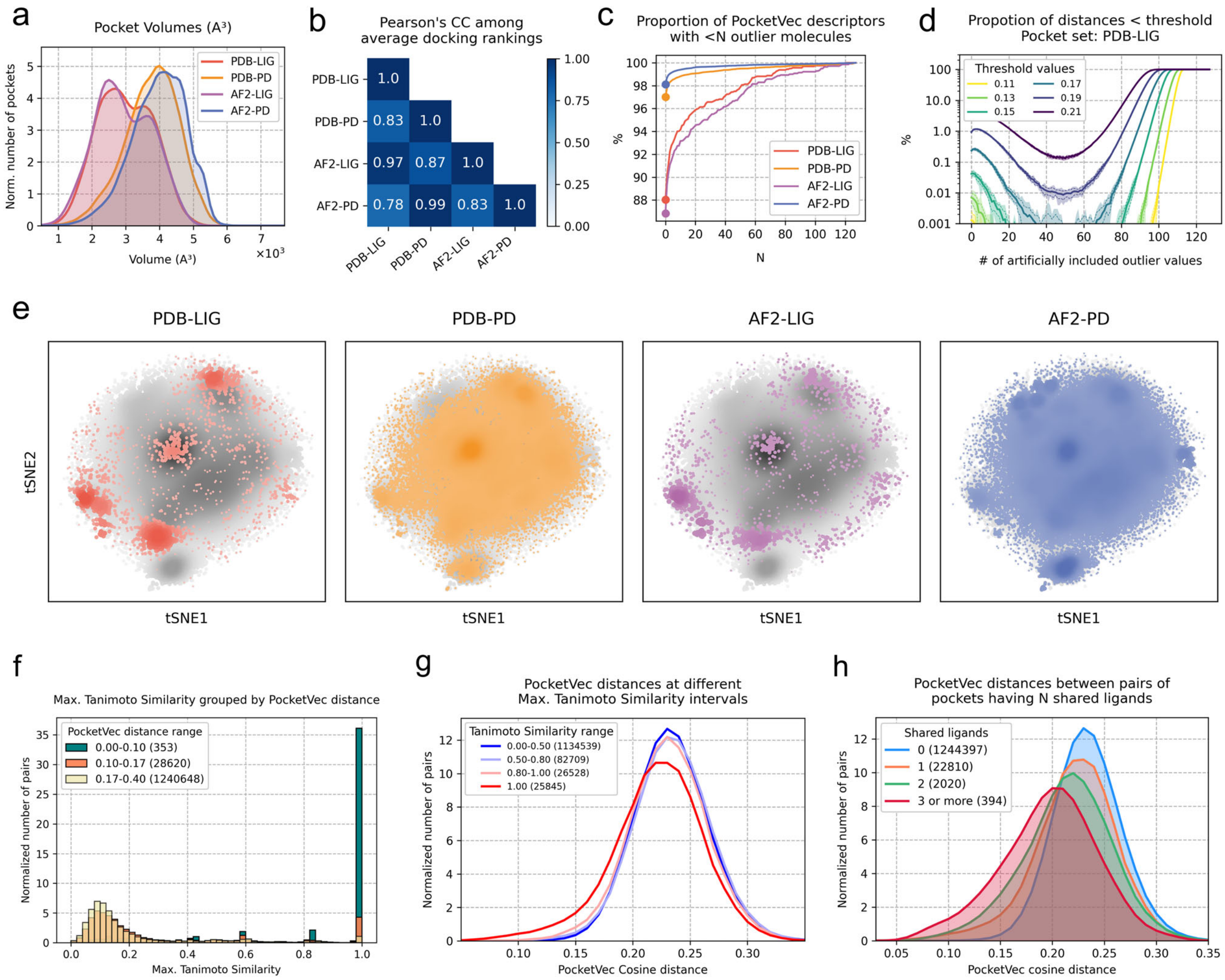
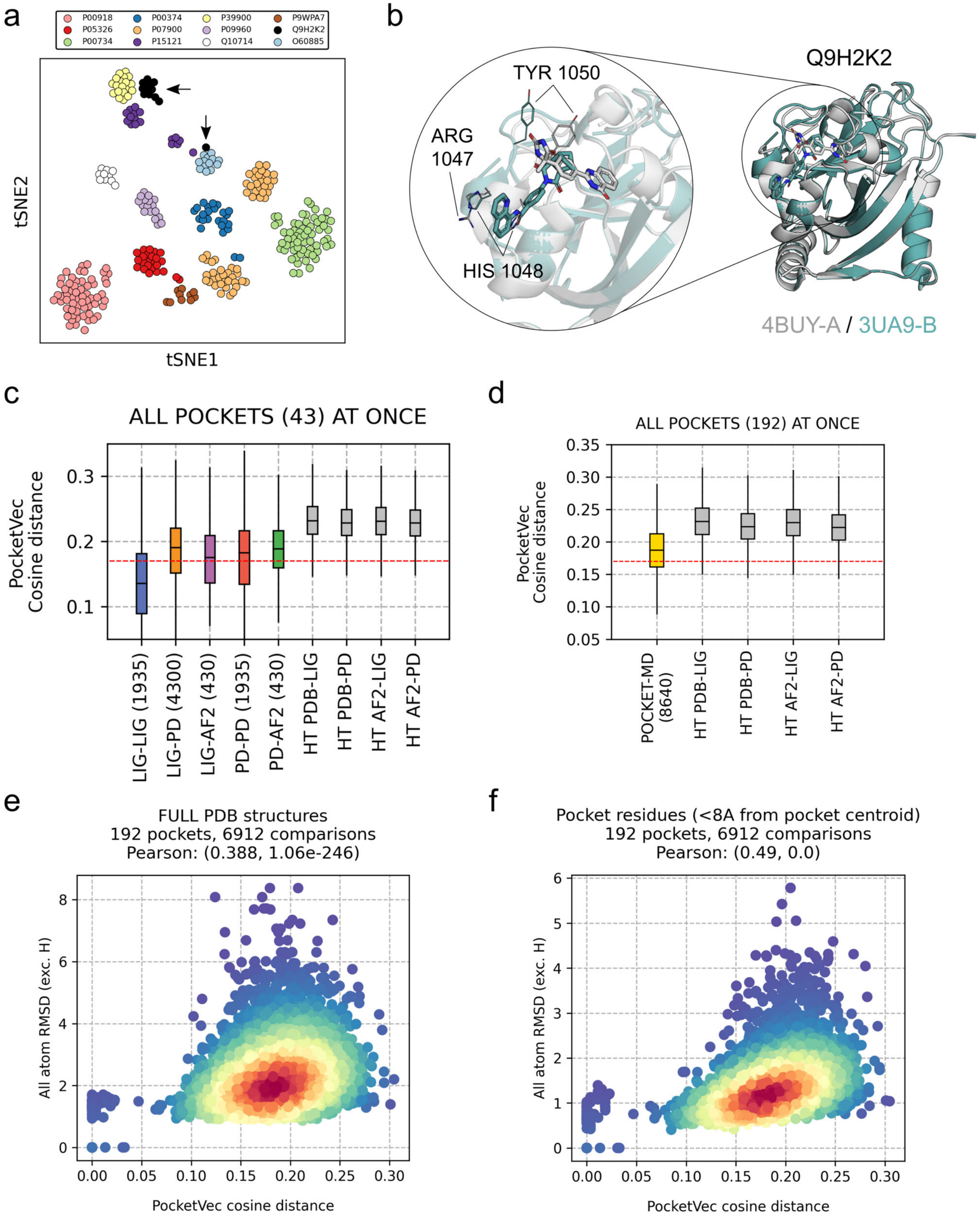


Fig. 3 | Characterization of human druggable pockets using PocketVec descriptors. a Normalized distribution (y-axis) of volumes (x-axis, x10^3 A^3) for each pocket set: PDB-LIG (1,604 pockets, red), PDB-PD (14,413 pockets, orange), AF2-LIG (1405 pockets, purple) and AF2-PD (32,302 pockets, blue). The average volumes are 2992.88 A^3, 3764.1 A^3, 2940.47 A^3 and 3972.66 A^3, respectively. Pocket volumes were directly calculated using the rDock CAVITY functionality. **b** Pearson's Correlation Coefficient between average docking rankings among all pocket sets. All *p*-values (10) are <10^{-25}. **c** For each pocket set, proportion of PocketVec descriptors (%, y-axis) having less than N (x-axis) outlier molecules (i.e. molecules with positive docking scores, please see *Methods: Post-docking analysis*). **d** Proportion of originally dissimilar pocket pairs (y-axis, distances >threshold) classified as similar (distances <threshold) due to the artificial insertion of a growing number of outlier values (x-axis) at different distance thresholds. To run the analysis, 100,000 PocketVec distances were randomly sampled (x10 times) from PDB-LIG. Data are presented as mean values (solid lines) +/- standard deviation percentages (dashed lines) together with minimum and maximum proportions among samples (shadowed areas). Results did not change among pocket sets (Fig. S23). **e** tSNE (t-distributed Stochastic Neighbor Embedding) representation of all PocketVec descriptors within each pocket set (PDB-LIG, PDB-PD, AF2-LIG, AF2-PD). Each pocket is represented by a single point, colored and sized by 2D density within the pocket set. Gray dots correspond to the background space: all PocketVec descriptors are considered at once, also colored and sized by 2D density. **f** Distributions (density, y-axis) of maximum Tanimoto Similarity among bound ligands (x-axis, bin width: 0.02) grouped by PocketVec distance ranges (0-0.10, 0.10-0.17 and 0.17-0.4). The number of pocket pairs per PocketVec distance range is specified in parenthesis. **g** Distributions (density, y-axis) of PocketVec cosine distances (x-axis) grouped by the maximum Tanimoto Similarity among bound ligands (0-0.5, 0.5-0.8, 0.8-1 and 1). The number of pocket pairs per maximum Tanimoto Similarity range is specified in parenthesis. **h** Distributions (density, y-axis) of PocketVec cosine distances (x-axis) grouped by the number of shared ligands between pockets (0, 1, 2 and 3 or more). The number of pocket pairs per number of shared ligands is specified in parenthesis. Source data are provided as a Source Data file.

PocketVec descriptors, leading to only ~0.039% of false positives (Fig. 3d and Fig. S23). In front of this strong robustness of the descriptors, in further analyses, we only discarded those very few PocketVec descriptors with more than 80 outlier molecules (10, 45, 15 and 43 descriptors from the PDB-LIG, PDB-PD, AF2-LIG and AF2-PD sets, respectively).

Influence of protein flexibility on PocketVec descriptors

Proteins are dynamic entities that may adopt various 3D conformations depending on multiple environmental factors (e.g. the presence of a ligand), as well as exhibiting a certain degree of flexibility in their side chains. Indeed, a single X-ray structure often does not represent the complete structural behavior of a protein, and we thus do not expect a single PocketVec descriptor to be a global representation of a pocket but instead a snapshot for each particular structure. To some extent, the ProSPECCTs benchmark set already assesses the sensitivity of PocketVec descriptors to the binding site definition (i.e. same pocket occupied by distinct ligands, P1 and P1.2) and the impact of protein flexibility (i.e. NMR-resolved structures, P2). In this context, we observed that, as expected, the variability among PocketVec descriptors from the same pocket was lower enough to be clearly distinguished from random pairs of pockets (Fig. 1d, AUROCs of 0.97,



1.00 and 0.96 for ProSPECCTs P1, P1.2 and P2, respectively). Besides, a 2D tSNE representation of PocketVec descriptors derived from the 326 structures in P1 showed a clustering pattern consistent with the 12 pockets represented (Fig. 4a). Interestingly, we observed that one of the structures (3UAB_B) of Poly-ADP polymerase tankyrase-2 (Q9H2K2) significantly deviated from the rest of the descriptors from the same pocket (protein) in the tSNE representation (black dots), and a visual inspection of its structure showed significant conformational changes in the side chains conforming the binding pocket with respect to the other ones (Fig. 4b), which translated into a higher PocketVec distance (>0.17) with the rest of the descriptors of the same pocket. In any case and, as briefly mentioned above, PocketVec descriptors for *holo* structures of the same pocket (ProSSPECTs P1 similar pairs) were fairly similar, with almost 80% of the pairs

Fig. 4 | Assessment of the effect of protein flexibility. a 2D tSNE representation of 326 PocketVec descriptors representing the conformational ensemble of 12 pockets binding distinct ligands (ProSPECCTs P1). Black arrows indicate the Poly-ADP polymerase tankyrase-2 (Q9H2K2) structures. **b** Structural superimposition of 4BUY_A (white) against 3UA9_B (teal) performed with TM-align. **c** PocketVec descriptors' variability caused by protein flexibility and conformational changes. Each boxplot indicates the distribution of PocketVec distances of the same pocket in i) *holo* PDB structures (LIG-LIG, $n = 1935$), ii) *holo* and *apo* PDB structures (LIG-PD, $n = 4300$), iii) *holo* PDB and AF2 structures (LIG-AF2, $n = 430$), iv) *apo* PDB structures (PD-PD, 1935), v) *apo* PDB and AF2 structures (PD-AF2, $n = 430$) and *holo* (LIG), *apo* (PD) and AF2 structures (AF2) against each of the 4 sets of PocketVec descriptors from our study (i.e. PDB-LIG, PDB-PD, AF2-LIG, AF2-PD. $n = 1{,}439{,}382$, $n = 12{,}974{,}304$, $n = 1{,}255{,}170$ and $n = 29{,}039{,}577$, respectively). Boxplots indicate median (middle line), 25th, 75th percentile (box), and max and min value within the 1.5 × 25th and 1.5 × 75th percentile range (whiskers). **d** PocketVec descriptors' variability caused by protein flexibility and conformational changes among MD-derived structures. POCKET-MD represents the distances between same-pocket MD-derived PocketVec descriptors ($n = 8640$) and all the other boxplots represent the distances between such MD-derived descriptors and each of the 4 sets already precomputed (i.e. PDB-LIG, PDB-PD, AF2-LIG, AF2-PD. $n = 3{,}060{,}480$, $n = 27{,}586{,}560$, $n = 2{,}668{,}800$ and $n = 61{,}745{,}280$, respectively). Boxplots indicate median (middle line), 25th, 75th percentile (box), and max and min value within the 1.5 × 25th and 1.5 × 75th percentile range (whiskers). **e** Correlation between PocketVec cosine distance (x-axis) and all-atom (excluding hydrogens) RMSD for the full PDB structure and **f** pocket residues only. The study involved 192 pockets (8 from PDB-LIG and 184 from PDB-PD) with MD-simulation data (3 frames per replica, 3 replicas). Both plots are colored by density (the redder the higher). Source data are provided as a Source Data file.

showing distances <0.17, while only 1,2% of the pairs showed distances <0.17 for dissimilar pocket pairs. When comparing *holo* structures to the corresponding *apo* structures or AF2 models of the same pocket, we found that the PocketVec descriptors showed higher variability than within *holo* structures, reflecting pocket structural heterogeneity (Fig. S24). However, the distributions of distances comparing *holo* with *apo* or AF2 structures were very similar, with 51% and 55% of pairs below the 0.17 threshold, respectively. To further complement this result, we performed an exhaustive evaluation of the behavior of PocketVec descriptors upon protein flexibility and conformational changes. In short, we kept those pockets from the PDB-LIG set for which we had enough *holo* and *apo* PDB structures available (≥10, see *Methods: Assessment of the effect of protein flexibility*) as well as a confident AF2 model, and generated PocketVec descriptors for all their structures. Overall, we collected 903 PocketVec descriptors corresponding to 43 unique pockets (10 *holo* structures, 10 *apo* structures and 1 AF2 model each). We then computed PocketVec distances between same-pocket *holo* PDB structures (LIG-LIG), same-pocket *holo* and *apo* PDB structures (LIG-PD), same-pocket *apo* PDB structures (PD-PD), same-pocket *holo* PDB and AF2 structures (LIG-AF2) and, finally, same-pocket *apo* PDB and AF2 structures (PD-AF2). Additionally, we compared *holo* (PDB), *apo* (PDB) and AF2 structures against our 4 sets of precompiled PocketVec descriptors (i.e. PDB-LIG, PDB-PD, AF2-LIG and AF2-PD) in order to contextualize the results with background distance distributions (Fig. 4c). We observed that PocketVec descriptors derived from *holo* PDB structures were rather robust and clearly distinguishable from random pocket pairs (LIG-LIG, AUROC of 0.90 against background distance distributions), in perfect agreement with the results obtained in ProSPECCTs P1 and P1.2. When we compared descriptors derived from *holo* and *apo* PDB structures, as expected, the intrinsic and increased variability of ligand-free structures was reflected in the PocketVec distances, which affected the ability to recapitulate pocket structures from the same protein (LIG-PD, AUROC of 0.77; PD-PD, AUROC of 0.80). Finally, we also observed that descriptors derived from *holo* structures tended to be more similar to those derived from AF2 models than to those from *apo* PDB structures (LIG-AF2, AUROC of 0.82). This observation aligns with several studies stating that AF2-predicted structures are comparable to the *apo* conformations and, perhaps, a bit closer to the *holo* conformations due to the nature of the training data (which includes both *holo* and *apo* structures)[69].

Additionally, we sampled MD trajectories for 192 pockets from the PDB-LIG and PDB-PD sets, considering 10 conformations per pocket, and then, for each pocket, we calculated all pairwise distances among their PocketVec descriptors (see Methods: *Assessment of the effect of protein flexibility*). As in the previous exercise, we also compared the obtained PocketVec descriptors with a random sample of precomputed descriptors from the PDB-LIG, PDB-PD, AF2-LIG and AF2-PD sets (Fig. 4d). Overall, we observed a similar trend than when comparing *apo* PDB structures: PocketVec descriptors showed variations responding to the flexibility of the proteins, while still recognizing different conformations from the same pocket (AUROC of 0.78). Finally, we analyzed the correlation between PocketVec distance among MD frames and the all-atom RMSD (excluding hydrogen atoms) considering both complete structures and only pocket residues (at <8 Å from the pocket centroid in any of the MD used frames). We found a weak correlation (Fig. 4e; Pearson CC 0.39) with the full structure RMSD that increased (Fig. 4f; Pearson CC 0.49) when exclusively considering pocket residues.

The presence of metal atoms is often required for many pockets to perform its native function (e.g. to act as catalytic centers or to stabilize protein structures). However, our study does not explicitly consider the presence of other compounds such as cofactors or metal atoms which, despite the limitations, it enables a fair and analogous characterization with different protein structures and pocket identification strategies where bound metal atoms are not available (e.g. PDB vs AF2 or LIG vs PD). Thus, finally, we also assessed the ability of our precomputed PocketVec descriptors to identify pocket similarities among metal-binding proteins. For this, after collecting all PDB-LIG and PDB-PD pockets with bound metal atoms, we compared the derived PocketVec descriptors with and without the metal atoms (*Methods: Assessment of the effect of metals binding proteins*). We observed that, although descriptors were obviously not identical, the vast majority of the identified metal-binding pockets had similar PocketVec descriptors with and without the explicit presence of the metal atoms (Mann-Whitney p-value $< 10^{-100}$; Fig. S25).

Overall, these results illustrate the capacity of PocketVec descriptors to capture pocket flexibility and protein conformational changes, revealing their sensitivity to changes in the pocket shapes as well as their feasibility of use with *holo* and *apo* PDB structures (including metal-binding proteins) and AF2 predicted models.

All-vs-all comparison of human druggable pockets

The vector-like nature of PocketVec descriptors makes them perfectly suited for fast comparisons by computing simple cosine distances, enabling comprehensive proteome-wide similarity searches. We first generated a *t*-distributed Stochastic Neighbor Embedding (t-SNE) representation for all PocketVec descriptors to facilitate a qualitative visualization of the characterized pocket space (Fig. 3e). This visualization underlined several interesting points, some of which have already been discussed in previous sections: (i) pockets defined by bound ligands (PDB-LIG and AF2-LIG) predominantly occupy well defined regions of the pocket space, (ii) pocket detection strategies (PDB-PD and AF2-PD) significantly contribute to expand the overall coverage of the human druggable pocket space, (iii) the use of AF2 models particularly bolsters this expansion. Perhaps more interestingly, the PDB-PD and AF2-PD maps highlight dense areas of the human pocket space for which we do not yet have any experimental structure with a bound chemical compound.

Then, we systematically evaluated the validity of the chemogenomic hypothesis behind the generation of PocketVec descriptors (i.e. similar pockets bind similar ligands). We compared all PDB-LIG pockets having PocketVec descriptors (1594 pockets, ~1.27 M comparisons) on the basis of the maximum Tanimoto similarity among their bound ligands using ECFPs (2048 bits and radius 2), and we grouped pocket pairs according to their cosine PocketVec distance (Fig. 3f). We found that, indeed, pocket pairs with small PocketVec distances (<0.10) typically showed higher maximum Tanimoto similarities (>0.85) among their ligands than pocket pairs with high PocketVec distance (>0.17, Fisher's exact test OR > 90 and p-value < 10^{-300}), thus supporting our underlying hypothesis. However, the fact that similar pockets bind similar ligands does not necessarily imply that similar ligands bind similar pockets. Indeed, we observed that, in general, PocketVec distances did not decrease substantially when increasing the maximum Tanimoto similarity among bound ligands (Fig. 3g). The only exception was for almost identical ligands (Tanimoto similarities=1), which showed a very subtle deviation towards smaller PocketVec distances (AUROC 0.59, true positives having Max. Tanimoto Similarity=1, true negatives having Max. Tanimoto Similarity in the [0, 0.5) range), in line with the behavior observed in ProSPECCTs P5 (AUROC = 0.64) and P5.2 (AUROC = 0.62). While this may seem somewhat counterintuitive, it serves as evidence to decipher the quantification of pocket similarity using PocketVec descriptors.

The fact that a single compound is consistently upranked for two pockets may indeed be a first evidence of pocket similarity but falls short to provide similar PocketVec descriptors if no other compound is ranked in a systematic manner. In practical terms, this translates into pockets that share a growing number of ligands being more likely to be similar from a PocketVec perspective, as highlighted by Shoichet and co-workers when they presented their similarity ensemble approach (SEA) to unveil protein remote relationships[57]. Reassuringly, we found that pockets that shared ligands in the PDB-LIG set tended to be more similar (i.e. smaller PocketVec distances) than pockets sharing no ligands (Fig. 3h). However, the deviation towards smaller distances for pockets sharing a single ligand was subtle (AUROC = 0.58, true positives sharing 1 ligand and true negatives sharing no ligands), but when 3 or more ligands were shared between pockets, the effect was already notable (AUROC = 0.75, true positives sharing 3 or more ligands and true negatives sharing no ligands). We also found that only 2.1% of pocket pairs sharing no ligands in the PDB-LIG set showed PocketVec distances <0.17, while this fraction increased to 26.1% when 3 or more ligands were shared between pockets. However, it is obvious that the lack of shared co-crystallized compounds between two pockets does not necessarily mean that they might not have common ligands. To overcome the limitation of potentially missing ligands, we used a pure computational strategy: we computed docking scores for the 128 standard lead-like molecules (see *Methods: Small molecule docking strategies, rigid docking*) against all PDB-LIG defined pockets and labeled them as active (the top 1% of docking scores) or inactive (Fig. S26a). We observed that the vast majority of the lead-like molecules (127 out of 128) were cataloged as active in, at least, one pocket (Fig. S26b), and almost 20,000 of the potential ~1,29 M PDB-LIG pocket pairs shared, at least, one active lead-like molecule (Fig. S26c). With this alternative approach, we confirmed the tendency observed using experimental data: the more shared ligands between pockets, the smaller their PocketVec distance (AUROC = 0.87 when comparing pockets sharing no ligands and those sharing 3 or more ligands; Fig. S26d).

Next, we explored the complementarity and added value of PocketVec descriptors with respect to more established strategies to compare protein families and druggable pockets, such as sequence and structure similarity. First, we sought to investigate the correlation between sequence, structure and PocketVec similarity (defined as 1-PocketVec distance) among pockets located in the same Pfam domains in the more comprehensive set of AF2-PD pockets (see *Methods: Systematic comparison of druggable pockets within domain families*). As expected, we found that the higher the sequential and structural similarity between compared pockets (assessed by sequence identity and Cα RMSD, respectively), the more similar they were according to PocketVec descriptors (Pearson CC of 0.55 and −0.35, respectively, p-value < 10^{-100} in both cases; Fig. 5a). Reassuringly, the observed correlations were also found when computed on AF2-LIG pockets (Pearson CC of 0.57 and −0.35 for sequence identity and Cα RMSD, respectively Fig. S27). However, there were indeed cases where the results did not align perfectly, underscoring the distinctive and complementary insights that PocketVec descriptors can provide beyond traditional sequential and structural analyses.

Additionally, we also ran an all-against-all pocket comparison within and across pocket sets (PDB-LIG, PDB-PD, AF2-LIG and AF2-PD), computing over 1.2 billion pocket comparisons. Interestingly, we found more than 3.5 million similar pockets in domains having low sequential and structural similarities (PocketVec distance <0.17; TM-score <0.35; Sequence identity <30%). For instance, we found similar pockets (PocketVec distance: 0.14, both pockets in the PDB-LIG set) in the CPSase_L_D2 ATP-binding domain (PF02786) of the Carbamoyl-phosphate synthase (P31327, positions 1088-1291) and in the NDK domain of the Nucleoside diphosphate kinase 3 (Q13232, positions 22-156), although they shared a sequence identity of only 28% and a poor structural similarity (TM-score=0.31, RMSD = 4.3 Å). Reassuringly, crystal structures confirmed that both pockets can bind ADP (PDB IDs: 5DOU and 1ZS6, respectively), which strengthened our observation that these pockets were indeed similar (Fig. 5b). The inventory of all similar pockets (PocketVec distance <0.17) together with structural and sequential comparisons at domain level are reported in our GitLab. On the other hand, our analyses also revealed more than 29k pocket pairs (out of a subsample of 11.1 million pairs having PocketVec distance >0.20) that, despite being similar in terms of sequence and structure, showed quite dissimilar pockets (PocketVec distance >0.20; TM-score >0.50; Sequence identity >40%). As an illustrative example, we identified different druggable pockets (PocketVec distance: 0.21, both pockets in the AF2-PD set) in the NIPSNAP domain (PF07978) of the Protein NipSnap homolog 3B (Q9BS92, positions 146–245) and in the NIPSNAP domain (PF07978) of the Protein NipSnap homolog 3 A (Q9UFN0, positions 146–245), although these domains had a sequence identity of 93% and also a very high level of structural similarity (Fig. 5c, TM-score = 0.98 and RMSD = 0.4 Å).

Relationship between PocketVec similarity and experimentally determined compound-target pairs

To further validate the ability of PocketVec descriptors to identify similar protein binding pockets, we explored the relationship between pocket similarity and experimentally determined compound-target pairs by assessing the number of shared compounds between proteins with different degrees of pocket similarity. We processed data on 836,654 compounds bound to 6933 protein targets from ChEMBL[75] and BindingDB[76], and we also collected all PocketVec descriptors from our study (i.e. PDB-LIG, PDB-PD; AF2-LIG and AF2-PD) and kept only those protein pairs for which there was, at least, one reported bound small molecule per protein (*Methods: Relationship between PocketVec similarity and experimentally determined compound-target pairs*). Overall, we evaluated 2,055,378 protein pairs on the bases of the number of shared compounds and the cosine distances between their PocketVec descriptors. Note that, since the experimental data reported compound-target pairs, without information on specific pockets, we always took the minimal PocketVec distance between each pair of proteins (i.e. the most similar pockets).

We first calculated the number of protein pairs within each range of PocketVec distances in an all-vs-all comparison of PocketVec descriptors, regardless of their set of origin (e.g. PDB-LIG, AF2-PD, etc) (Fig. 6a). Then, we checked the distribution of shared compounds

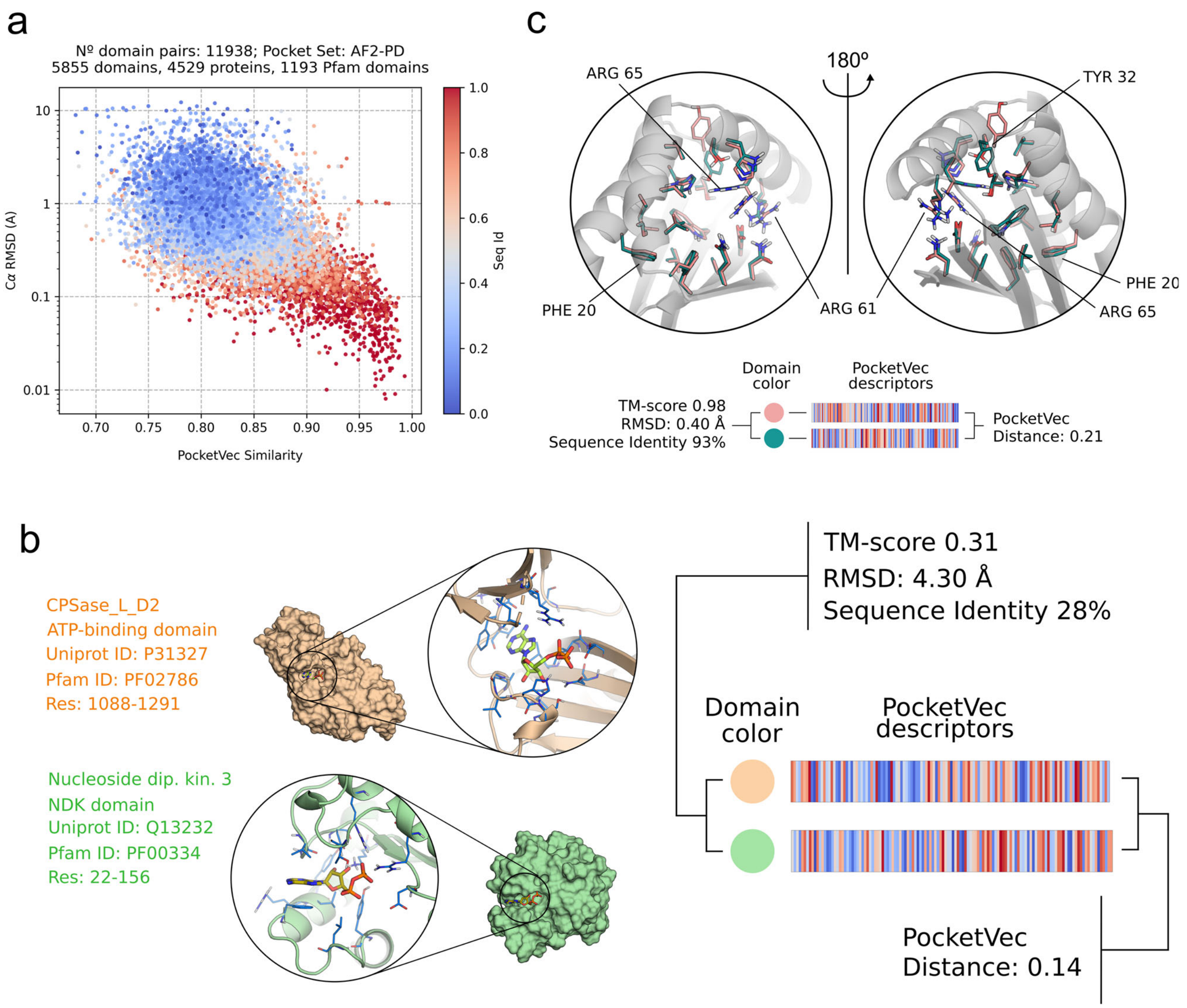


Fig. 5 | Using PocketVec descriptors to assess proteome-wide pocket similarity: achieving otherwise unattainable insights. a Correlation between PocketVec similarity (x-axis, defined as 1-PocketVec distance), structural similarity (y-axis, Cα RMSD) and sequence identity (color) among pockets located at the same Pfam domains (max. 10) in the AF2-PD pocket set. Pearson CC between PocketVec similarity and sequence identity: 0.55 (p-value ~0). Pearson CC between PocketVec similarity and RMSD: −0.35 (p-value ~0). **b** Similar pockets found in dissimilar domains. Pockets were found in the CPSase_L_D2 ATP-binding domain (PF02786, top structure, wheat color) of the Carbamoyl-phosphate synthase (P31327, positions 1088–1291. PDB ID: 5DOU) and in the NDK domain (PF00334, bottom structure, pale green color) of the Nucleoside diphosphate kinase 3 (Q13232, positions 22-156. PDB ID: 1ZS6). Pockets (both in the PDB-LIG set) have a PocketVec distance of 0.14 (below the established threshold of 0.17, please see *PocketVec performance on the ProSPECCTs benchmark sets*) although domains share a sequence identity of only 28% and a poor structural similarity (TM-score=0.31, RMSD = 4.3 Å). **c** Dissimilar pockets found in similar domains. Pockets were found in the NIPSNAP domain (PF07978) of the Protein NipSnap homolog 3B (Q9BS92, positions 146-245, AF2 model, pink residues) and in the NIPSNAP domain (PF07978) of the Protein NipSnap homolog 3 A (Q9UFN0, positions 146-245, AF2 model, green residues). The former is used as the reference structure (gray cartoon). Pockets (both in the AF2-PD set) have a PocketVec distance of 0.21 (above the established threshold of 0.17, please see *PocketVec performance on the ProSPECCTs benchmark sets*) although domains have a sequence identity of 93% and also a very high level of structural similarity (TM-score=0.98 and RMSD = 0.4 Å). Source data are provided as a Source Data file.

between each protein pair plotted against their minimal PocketVec distance. We observed that, the lower the minimal PocketVec distance between two proteins, the higher the number of compounds binding them both and, for PocketVec distances ≥ 0.2, there were almost no shared compounds for any protein pair (Fig. 6b). Further, we calculated the odds ratio considering a varying number of shared compounds for PocketVec distances ≤0.17, ≤0.10 and ≤0.05. We found that, indeed, there was a clear enrichment of protein pairs sharing compounds among protein pairs with similar pockets. (Fig. 6c). For instance, we found ~2, ~50 and ~200 fold enrichments of protein pairs sharing ≥ 5 compounds in protein pairs with minimal PocketVec cosine distances below 0.17, 0.10 and 0.05, respectively. And these enrichments went up to ~5, ~300 and ~700 if we considered protein pairs sharing ≥ 20 compounds. Finally, to overcome potential biases due to the comparison of PocketVec descriptors from the same origin (e.g. PDB-LIG, AF2-PD, etc), we repeated the analysis considering only comparisons between the PDB-LIG set (i.e. derived from experimental structures with co-crystallized compounds) and the other sets, observing a very similar result (Fig. S28a, b, c).

Additionally, we checked the robustness of our results on a recent dataset where Winter and co-workers comprehensively tested the potential binding of 407 fragment compounds on 5951 proteins. After applying the filtering criteria to the raw data defined by the authors, we analyzed 301 fragment compounds and 525 proteins for which we had PocketVec descriptors (Methods: *Relationship between PocketVec similarity and experimentally determined compound-target pairs*). We

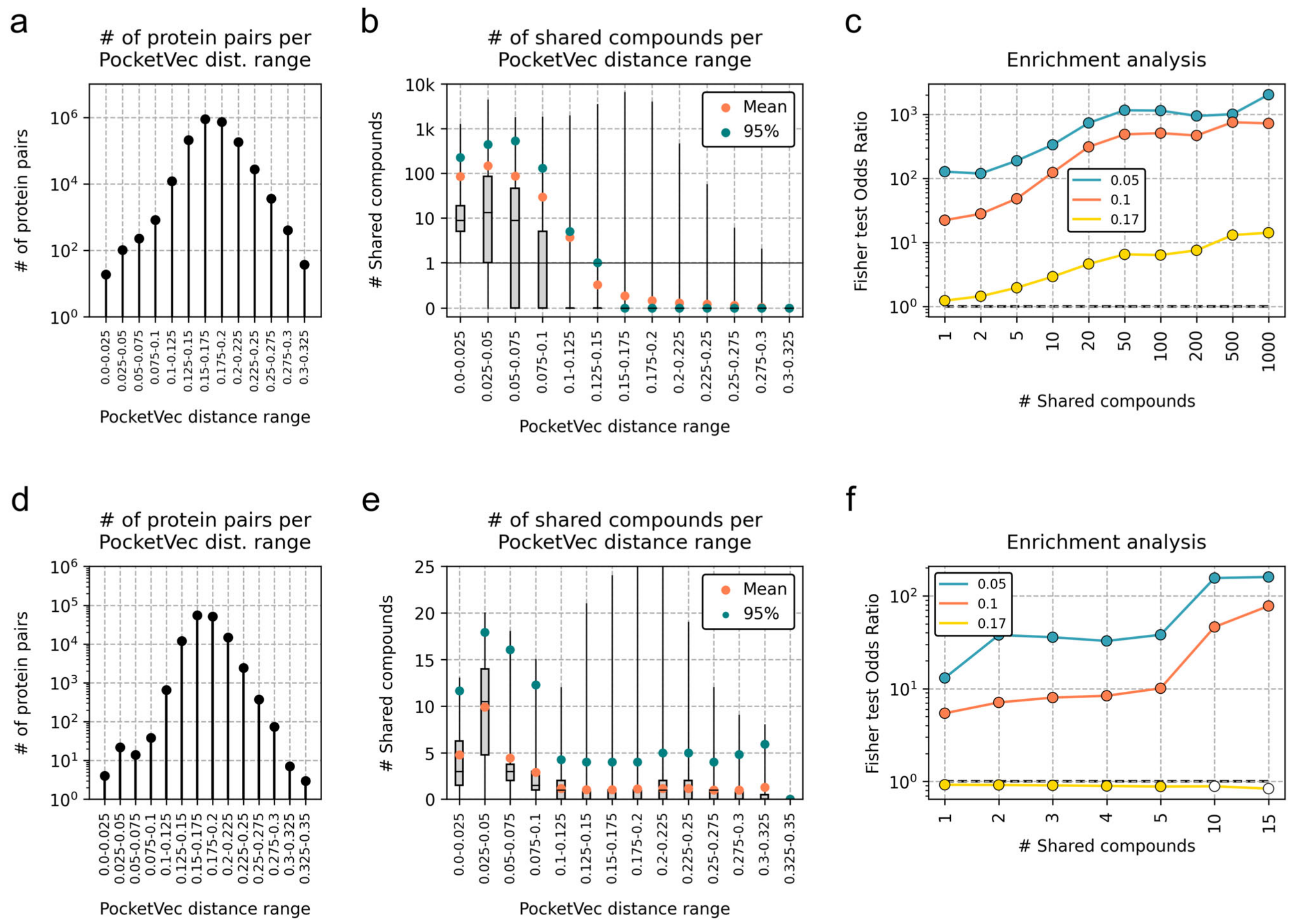


Fig. 6 | Relationship between PocketVec similarity and experimentally determined compound-target pairs. All vs All (PDB-LIG, PDB-PD, AF2-LIG and AF2-PD) comparison between PocketVec distance and number of shared compounds among proteins having experimental binding data in ChEMBL and BindingDB (**a**, **b**, **c**) and[105] (**d**, **e**, **f**). **a**, **d** Number of protein pairs (y-axis) at each PocketVec distance range (x-axis). **b**, **e** Number of shared compounds for all pairs having a PocketVec distance in the specified distance range. The number of points in each boxplot is specified in Fig 6a,d. Orange dots indicate the average value while green dots indicate the lower bound for the top 5% of the pairs. Boxplots indicate median (middle line), 25th, 75th percentile (box), and maximum and minimum values (whiskers). **c**, **f** Evolution of Fisher test (two-sided) Odds Ratio (y-axis) for an increasing number of shared compounds (x-axis) using an increasing value of PocketVec distance cut-off (0.05, 0.10 and 0.17). Colored dots indicate p-value < 0.001, gray dots indicate p-value < 0.05 and white dots indicate p-value> 0.05. Source data are provided as a Source Data file.

repeated the analyses described above finding that, despite the much lower number of instances (from ~2.1 M protein pairs to ~140 k), protein pairs with similar pockets tended to share more fragments (Fig. 6d, e), which also translated in significant enrichments (Fig. 6f). In this case, we found no enrichment of common compounds for PocketVec distances ≤0.17, which was expected due to the smaller nature of the fragments (translating in lower specificity). However, we observed enrichments of protein pairs sharing ≥ 3 compounds of ~8 and ~30 fold for PocketVec distances ≤0.10 and ≤0.05, respectively. For completeness, we also ran the same analyses considering only comparisons between the PDB-LIG and the other PocketVec descriptor sets (Fig. S28d, e, f). However, although we still observed a similar trend, the counts were too low to extract any significant conclusion.

Overall, these results show that, indeed, there is a clear relationship between pocket similarities, as defined by low PocketVec distances, and the probability of those proteins to experimentally bind the same compounds, further validating our approach.

Identifying kinases with similar inhibition profiles through PocketVec descriptors

Protein kinases have long been a cornerstone of drug discovery efforts primarily due to their role as oncogenic targets[77]. Since the breakthrough FDA-approval of imatinib in 2001, dozens of small molecules (72 as of November 2022[78]) have been approved as therapeutic kinase inhibitors for clinical use. However, the design of selective inhibitors is a challenging task due to the highly conserved ATP-binding pocket shared among human kinases. Indeed, many kinase inhibitors show high promiscuity (i.e. they bind to many kinases), while others are rather selective[79,80]. This same variability is also apparent from the kinase perspective: some bind to many inhibitors while others are very selective[81]. Our characterization of druggable pockets in human proteins showed 1286 potential binding sites within protein kinase domains. More specifically, we derived PocketVec descriptors for 229, 404, 195 and 458 pockets in the PDB-LIG, PDB-PD, AF2-LIG and AF2-PD sets, respectively. Thus, we set up to explore a potential correlation between the pockets in the different kinases and their experimentally determined small molecule inhibition profiles.

We first collected kinase-inhibitor pairs identified through systematic chemical proteomics approaches by Kuester and collaborators, where they tested interactions between 520 kinases and 243 inhibitors (2017 set)[82], and between 318 kinases and 1183 inhibitors (2023 set)[83]. We used a standard activity cutoff of 30 nM to define binary kinase-inhibitor matrices, as recommended in Pharos 17 (http://pharos.nih.gov). In the 2017 set, we found interactions involving 111 kinases and 94 inhibitors (Fig. 7a), with 43 kinases being inhibited by a single compound, and 18 of them interacting with 5 or more inhibitors

(Fig. S29a). In the 2023 set, we identified interactions comprising 73 kinases and 164 inhibitors, where 19 kinases interacted with 1 inhibitor and 27 with 5 or more (Fig. S29b). We then compared kinases on the basis of their binarized inhibition profiles (Jaccard similarity, Fig. 7d, j, upper triangular matrices), and we observed that a relatively small fraction of kinase pairs shared at least 1 inhibitor (12.3% and 9.7% for the 2017 and 2023 sets, respectively). We also compared the kinases using PocketVec descriptors (Fig. 7d, j, lower triangular matrices), finding that the fraction of kinase pairs showing similar druggable pockets (i.e. PocketVec distances <0.17) was 40.5% and 41.4%, respectively. As expected, these fractions were significantly higher than the figure obtained when comparing random sets of pocket descriptors (~2%. Fisher's exact test, OR > 30, p-value ~0), since all kinases contain at least one similar ATP-binding pocket. Reassuringly, as observed in the general analysis of human druggable pockets (Fig. 3h), we found that the higher the number of common inhibitors between pairs of kinases, the more similar their pockets were according to PocketVec descriptors (Fig. 7b, h). However, even when no shared inhibitor was found for a pair of kinases, as expected, the similarity between their ATP-binding pockets was quite remarkable, suggesting that the PocketVec distance threshold should be lowered to disentangle drug promiscuity.

Next, we parsed the general inhibition matrices to derive inhibition profiles for each kinase (i.e. a vector describing whether it does or does not interact with every tested inhibitor), and we assessed the coherence between PocketVec distances and the similarity between experimentally determined inhibition profiles. Indeed, kinase pairs with low PocketVec distances (<0.17) showed a significant enrichment for similar inhibition profiles (Fisher's exact test, OR = 3.2, p-value < 0.0003 in the 2017 set, and OR = 4.1, p-value < 0.003 in the 2023 set for a Jaccard similarity >0.5). These enrichments were even more pronounced when we applied more stringent distance thresholds (PocketVec distance <0.10) with OR > 118 and OR > 125 (p-values < 10^{-10}) for the 2017 and 2023 sets, respectively (Fig. 7c, i).

We then compared our results using PocketVec descriptors to more classical structure and sequence similarity analyses (see *Methods: Comparison of kinase inhibition profiles with PocketVec descriptors, sequence and structure similarity measurements*). As expected, we found that structurally similar proteins (Max. TM-score >0.85) were also moderately enriched in similar inhibition profiles (OR = 2.4, p-value < 0.05; OR = 1.7, p-value > 0.05), as were kinases sharing >35% sequence identity (OR = 4.8, p-value < 10^{-6}; OR = 5.5, p-value < 10^{-4}). Finally, we studied the relationship between structure, sequence, PocketVec and inhibition profile similarity between kinase pairs. Consistently with our global analysis of the human druggable pockets (Fig. 5a), higher structure and sequence similarities consistently led to lower PocketVec distances in both the 2017 (Fig. 7e, f) and the 2023 (Fig. 7k, l) sets.

Despite the coherence of the results of the three metrics, interestingly, we found 9 kinase pairs (4 in the 2017 and 5 in the 2023 sets) with clear inhibition profile similarities that only PocketVec descriptors could pick. On the other hand, our analyses also revealed 8 cases (5 in the 2017 and 3 in the 2023 sets) where PocketVec descriptors fell short to detect similarities that could be recovered by sequence or structure comparisons alone. Overall, these results emphasize the coherence and complementarity between strategies, highlighting the potential of PocketVec descriptors to identify otherwise unattainable similarities in experimental inhibition profiles among protein kinases.

Discussion

We have presented PocketVec, an approach to generate vector-like protein pocket descriptors based on inverse docking and the chemogenomics principle that similar pockets bind similar ligands. A thorough assessment of its performance ranks it among the best available methodologies to characterize and compare protein druggable pockets, while overcoming some important limitations. We have also systematically searched for druggable pockets in the folded human proteome, using experimentally determined protein structures and AF2 models, identifying over 32,000 binding sites in more than 20,000 protein domains. We then derived PocketVec descriptors for each small molecule binding site and took advantage of their vector-like format to run an all-against-all pocket similarity search, exploring over 1.2 billion pairwise comparisons. Besides, we provide precomputed descriptors for every identified human pocket together with the annotated Python code to generate descriptors for any pocket of interest. We found that PocketVec descriptors are complementary to other, more classical, search strategies, enabling the identification of pocket similarities not revealed by structure- or sequence-based comparisons. As illustrative examples of applicability, we unveiled a clear relationship between pocket similarities, as defined by low PocketVec distances, and the probability of those pockets to bind the same compounds, as experimentally detected. Moreover, a systematic comparison of druggable pockets in protein kinases showed that kinase pairs with similar PocketVec descriptors also exhibited similar experimentally determined inhibition profiles.

There have been recent attempts to identify druggable pockets at the proteome level (e.g[84–87].). However, our distinct ligand-centric methodological approach, the accuracy of our descriptors and the systematic and exhaustive identification and characterization of binding pockets enabled the analyses of the effects of protein structural variation on pocket definition and small molecule binding at a remarkably broad scale. Besides, the comprehensive list of human-derived pocket descriptors will become a valuable resource for the bio and chemoinformatics communities.

This generation of descriptors has been primarily designed for global analyses, such as the comprehensive characterization of all human druggable pockets. Indeed, our analyses have revealed dense clusters of similar pockets in distinct proteins for which no inhibitor has yet been co-crystalized, opening the door to strategies to prioritize the development of chemical probes to cover the druggable space[88]. Moreover, our initial descriptors can be easily adapted to cater to specific tasks (i.e. exploring substrate specificity in a given protein family) by refining the selection of predefined lead-like molecules used or fine-tuning the similarity cutoff, thereby enhancing their performance. Of special interest are the anticipation of undesired off-targets as well as the guidance of rational polypharmacology, where single univalent molecules could be designed to target two proteins simultaneously, provided that their druggable pockets are similar enough[33]. However, the main impact is likely to come from proteochemometric approaches, where a combination of ligand and target descriptors are used to train machine learning models[13]. It has been shown that structure-based descriptors of the targets are often superior to distinguish drug selectivity, although the sequence-based ones are often used when key protein structural details are lacking[39]. The generation of accurate descriptors derived for not yet described pockets in AF2 protein models overpasses this limitation, and open up many possibilities. We envisage a scenario where small molecule and pocket descriptors combined are used to train AI-based generative models (e.g[89,90]. to design chemical entities that bind each protein druggable cavity. Indeed, the estimated space of 10^{33} synthetically accessible drug-like molecules is mostly unexplored and represents a reservoir of potentially bioactive compounds[91]. Deep learning strategies have successfully designed antibiotic scaffolds[92] and placed 15 AI-designed drugs in clinical trials, including first-in-class molecules against several targets[93]. Overall, accurate descriptors of druggable pockets might serve as a cornerstone for the development of generative AI approaches in drug discovery, offering opportunities to expedite the design of a chemical toolbox to probe biology and, ultimately, to therapeutics.

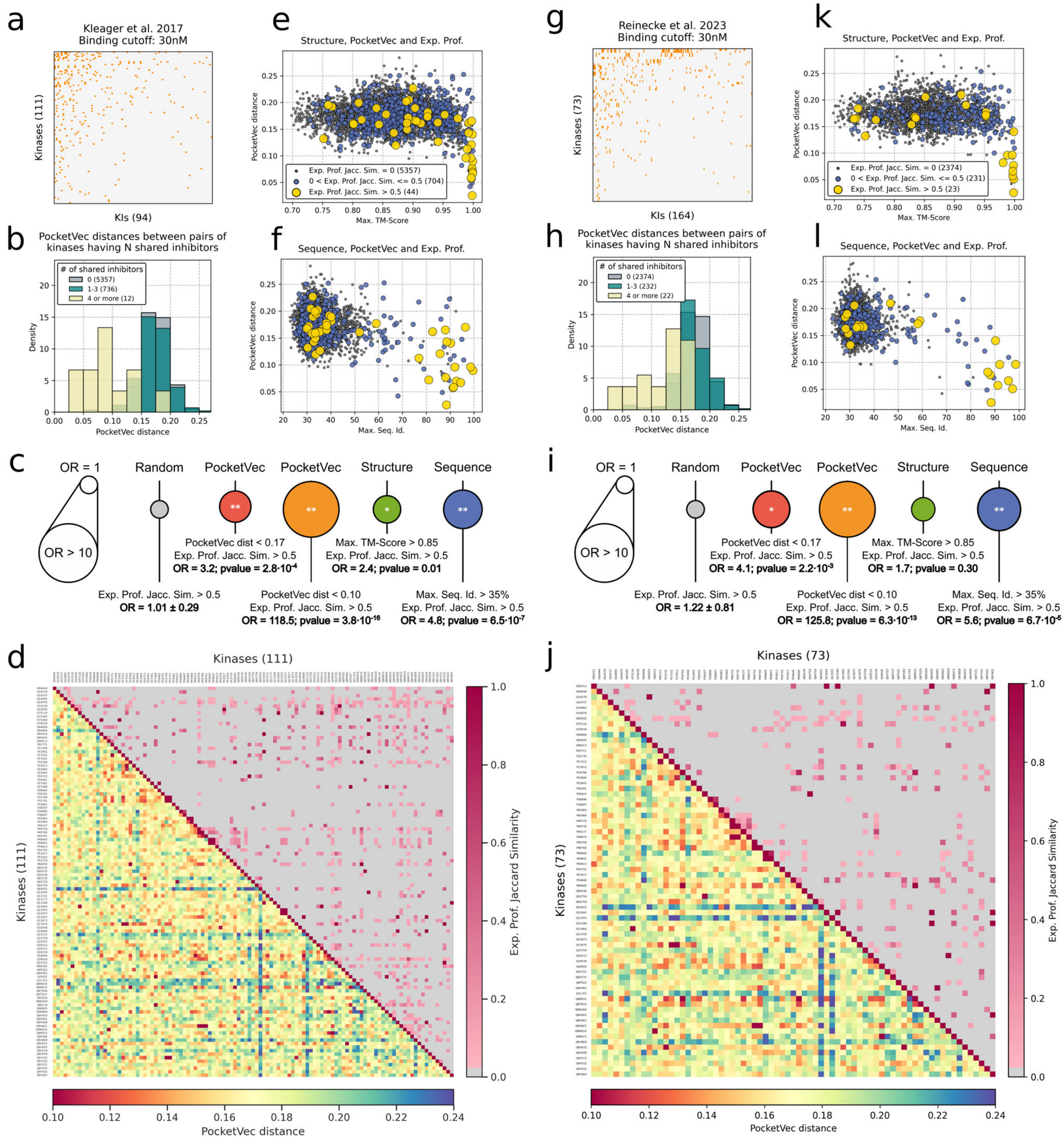


Fig. 7 | Correlation between inhibition profiles and PocketVec descriptors. All the analyses have been performed on the data obtained from Kleager et al.[82] (right panels) and Reinecke et al.[83] (left panels). **a**, **g** Inhibition matrix between protein kinases, and small molecule kinase inhibitors binarized at 30 nM. Both kinases and inhibitors are sorted by the number of active inhibition events. Orange dots indicate inhibition and white dots indicate no inhibition. **b**, **h** Distributions of PocketVec distances grouped by the number of shared inhibitors between kinases (0, 1–3 and 4 or more). The number of kinase pairs per number of shared inhibitors is specified in parenthesis. **c**, **i** Enrichments (Fisher's exact test, two-sided) in similar inhibition profiles (Jaccard Similarity >0.5) for those kinase pairs being similar in terms of PocketVec distance (red: <0.17, orange: <0.10), structural similarity (TM-score >0.85) and sequence identity (>35%). For comparison, the results obtained with randomly selected kinase pairs (gray) are also included. Circle areas are proportional to the corresponding ORs and p-values are specified in the center with the following format: *p-value < 0.05, **p-value < 0.001. **d**, **j** Pairwise kinase comparisons. Rows and columns correspond to alphabetically sorted kinases (by Uniprot ID). Upper triangular matrices: kinases are compared on the basis of their experimentally determined inhibition profiles. Each square represents the Jaccard similarity between the inhibition profiles of two targets: the higher the Jaccard similarity, the more similar the corresponding inhibition profiles. Lower triangular matrices: kinases are compared on the basis of their PocketVec descriptors (employing the minimum distance among all PocketVec descriptors in the Protein Kinase Domains). The color of each square indicates the minimum PocketVec distance between two targets: the lower the PocketVec distance (red), the more similar the kinases are at pocket level according to our descriptors. **e**, **k** Relationship between structural similarity (x-axis, Max. TM-score) and PocketVec distances (y-axis) between pairs of protein kinases. Each point represents a kinase pair and is colored and sized in terms of the similarity between experimentally determined inhibition profiles. **f**, **l** Relationship between sequence similarity (x-axis, Max. Seq. Id.) and PocketVec distances (y-axis) between pairs of protein kinases. Each point represents a kinase pair and is colored and sized in terms of the similarity between experimentally determined inhibition profiles. Source data are provided as a Source Data file.

Methods

Selection of compound sets

Fragments: we downloaded the Glide[60,61] diverse fragment dataset from the Schrodinger website (https://www.schrodinger.com) in November 2020. This collection of compounds is composed of 667 molecules having molecular weights the 50–200 g·mol^{-1} range (Fig. S2).

Lead-like molecules (LLM): we retrieved a set of 650 k lead-like molecules from MOE v2019.01 (Chemical Computing Group, Montreal, Canada). We then performed a k-means clustering using TAT fingerprints, setting the number of clusters to 1000, and selected the corresponding 1000 molecules closest to each cluster centroid to build the library. These compounds exhibit molecular weights in the 200-450 g·mol^{-1} range (Fig. S2).

Small molecule docking strategies

Rigid docking: we performed rigid docking calculations using rDock[62] (downloaded on July 2021 from https://github.com/CBDD/rDock). We prepared the protein structures using MOE v2019.01 (Chemical Computing Group, Montreal, Canada), Biopython[94] and the structure checking utility from BioExcel Building Blocks[95]. We ran all docking calculations using standard parameters and scoring functions. The binding site box was built around the pocket centroid (ligand centroid or detected pocket centroid) with a radius of 12 Å (*RbtLigandSiteMapper* option) and the number of runs was set to 25. Finally, we set the *DIHEDRAL_MODE* to *FIXED* (rigid docking). The considered score for each molecule was the minimum value of SCORE.INTER.

Flexible docking: we used SMINA[63] (downloaded on November 2020 from https://sourceforge.net/projects/smina/) for the flexible docking calculations. We prepared the protein structures using Reduce[96], OpenBabel[97], Biopython[94] and the structure checking utility from BioExcel Building Blocks[95]. We ran all the calculations using standard parameters and scoring functions for flexible docking. The binding site box was automatically derived from the position of the bound ligand (*autobox_ligand* parameter). The considered score for each molecule was the minimum value of *minimized_affinity*.

Post-docking analysis

For each pocket under evaluation, docking scores were stored in a one-dimensional NumPy array[98] and then translated into rankings using SciPy[99] (*rankdata* function, method *max*). In this way, the molecule with the lowest docking score was assigned the top ranking (1st), while the one with the highest docking score was ranked as the *N*th (being *N* the total number of tested molecules; $N = 128$ in the standard PocketVec pipeline). It is important to note that molecules yielding positive docking scores (e.g. due to structural clashes with the protein) were not explicitly considered and their corresponding rankings were set to an outlier value (e.g. 129). The rationale behind this procedure was that such outlier molecules were indeed informative (i.e. the pocket was too small to fit them) but needed to be distinguished from binders having poor (but negative) docking scores (i.e. weak binders). Specifically, docking scores in the range 0–50, 50–100 and >100 were translated into N + 1, N + 2 and N + 3 rankings, respectively (129, 130, 131 in the standard PocketVec pipeline).

Benchmark set

A good strategy to identify the best combination of small molecules and docking methods to develop pocket descriptors, and to assess their validity, is to see if they can faithfully capture reported similarities between small molecule binding pockets. To this end, we used ProSPECCTs[41], a collection of 10 datasets composed of protein-ligand binding site pairs classified as similar or dissimilar according to specific criteria (downloaded in July 2021 from http://www.ewit.ccb.tu-dortmund.de/ag-koch/prospeccts/). P1 (P1.2) includes 326 (45) protein-ligand complexes involving 12 (12) different proteins, and it is meant to study the sensitivity of pocket comparison tools to the binding site definition by comparing proteins having identical sequences with chemically distinct (similar) ligands located at the same site. P2 comprises 17 PDB files resolved by NMRs, containing a total of 329 different models, and was designed to assess the impact of protein flexibility in pocket comparisons. P3 and P4 include a variable number (1 to 5) of randomly added artificial mutations in the P1 proteins leading to changes in the physicochemical (P3) and physicochemical and shape (P4) properties of the protein binding site. For the sake of simplicity and coherence with reported performances, we have only considered structures with 5 mutations (representing 326 out of 1630 mutated structures). P5 (P5.2) was designed to detect pairs of unrelated proteins binding to identical or similar ligands, and consists of 80 (100, including phosphate binding sites) protein-ligand complexes[100]. P6 (P6.2) is intended to evaluate the identification of distant relationships between pockets binding to identical ligands but having variable pocket environments[101], and it includes 115 protein structures excluding (including) cofactors. We did not use P6 or P6.2 to benchmark our methodology since all pocket pairs are bound to identical or highly similar ligands, and the similar/dissimilar classification is done considering fine details of protein-ligand binding, such as the involved ligand functional groups. Thus, this set is not appropriate to guide and assess the developments of our pocket descriptors. Finally, P7 was retrieved from published successful applications of pocket similarity studies in a diverse set of proteins, and it contains 1151 protein structures. A detailed overview of all ProSPECCTs datasets is presented in Fig. S3.

Entropy measurements

For each molecule within each ProSPECCTs dataset, rankings were first binned into 100 different groups (bins) in order to discretize a variable (rankings) that, in practical terms, was continuous (since ranking range was usually higher than the number of considered structures). Shannon's Entropy was then calculated using such binned data (SciPy[99], *entropy* function, base 2).

Domain-based characterization of the human druggable pockets

We searched all human protein identifiers from UniProt (July 2022, *organism_id 9606* and *reviewed* set to *true*), retrieving a total of 20,386 unique human proteins[67]. Then, for each human protein, we retrieved all Pfam domains[68], considering only those entities labeled as 'domain' (e.g. we did not include 'repeats'). Overall, we found 28,044 domains (2704 unique Pfam domains) in 11,242 human proteins (Fig. 2a and Fig. S15).

To structurally annotate these domains, we used two different strategies. On the one hand, we looked for experimentally determined structures searching the PDB[2]. For each human domain, we gathered all PDB chains showing a structural coverage of the domain ≥80% using the localpdb package[102] (PDB version 2022.02.25). We identified at least one PDB structure for 7774 domains (1839 unique Pfam domains in 4726 proteins), processing all PDB files and removing those regions outside the domains under study. Additionally, we downloaded all predicted structures for Homo Sapiens from AlphaFold DB (https://alphafold.ebi.ac.uk/download#proteomes-section, proteins having <2700 amino acids, August 2022), processed all files and removed those regions that did not match the domains under study, leaving predicted structures for 25,589 domains (2671 unique Pfam domains in 11,022 proteins).

To identify druggable pockets, we also followed two complementary strategies.

Ligand-based. In the ligand-based pocket definition, we identified all the PDB structures (chains) corresponding to human protein domains that contained small molecules (HET PDB codes) cocrystallized with the domains of interest that fulfilled the following criteria: i) being annotated as ligands in PDBSUM data, ii) not being one of the 20 naturally occurring amino acids, iii) having a number of carbon atoms >6, to filter out solvent molecules and crystallography-related species and vi) having solvent accessibility ≤0.4 or buriedness ≥15. We defined solvent accessibility as the ratio between the ligand solvent-accessible surface area (SASA) in the bound state and the ligand SASA in the free state. SASA values were calculated with RDKit. Additionally, we defined buriedness as the number of protein residues having a distance below 8 Å to the ligand centroid, and we calculated these values using Biopython[94]. The cut-off values for accessibility (0.4) and buriedness (15) were set upon visual inspection of many bound ligands. We considered both parameters in order to recover as many interesting cases as possible: accessibility values usually underestimated pockets defined by small ligands while buriedness values often underrated large pockets. By applying all these conditions, we made sure to consider only pharmacologically relevant small molecules while filtering out inorganic chemicals, short peptides and crystallization additives such as ethylene glycol (EDO), glycerol (GOL), polyethylene glycol (PEG) and DMSO (DMS), among many others. In total, we found at least one PDB structure containing a ligand fulfilling all conditions for 1279 domains (363 unique Pfam domains in 1205 proteins). For 503 of these, we only found a single ligand, whereas for 254 of them we could find 10 or more ligands (Fig. S16), including a variety of metabolic nucleotides such as ADP (found in 114 domains, Fig. S30) or GDP (found in 102 domains, Fig. S30).

To compile the list of unique ligand-defined pockets we followed the procedure shown in Fig. 2a. First, we chose a reference PDB structure for each protein domain where we considered the structural coverage of the domain and the resolution of the crystal structure. We then superimposed all domain structures with the corresponding bound ligands onto their reference using TM-align[103].

To define a final set of ligand-based druggable pockets per domain, we used a single-linkage clustering technique, merging into a single pocket all those ligands whose centroids were at a distance ≤5 Å while maintaining the maximum distance between the global centroid of the cluster and the centroids of the individual compounds ≤18 Å. We considered the final global cluster centroids as the pocket centroids. Overall, we found 1604 ligand-defined pockets in 1279 protein domains (363 unique Pfam domains in 1205 proteins). We named this set of pockets PDB-LIG.

We then superimposed the reference PDB structure of the previous domains to their AF2 predicted counterparts by means of TM-align[103], and transferred the location of the identified PDB-LIG pockets. We only considered those pockets having a pLDDT value > 70 for all the residues at a distance ≤8 Å from the pocket centroid. Overall, we identified 1405 pockets in 1131 domains (339 unique Pfam domains in 1074 proteins), and named this set of pockets AF2-LIG.

Pocket-detection. As a complementary strategy, and to increase the overall coverage of human druggable pockets, we attempted a de novo identification of pockets. To establish a standardized protocol to predict them, we assessed the accuracy of different methods when identifying the PDB-LIG pockets defined above. In brief, we first removed bound ligands from reference *holo* structures (1279 PDB structures, one per domain) and used Fpocket[73] and P2rank[74], two state-of-the-art methods, to detect pockets in ligand-free domain structures. Additionally, we also used Prank[74], a functionality of P2rank aimed at rescoring the pockets predicted by Fpocket. In this way, we benchmarked three different strategies to detect and score domain binding sites. We considered that a predicted pocket and a ligand-defined pocket matched if the distance between their centroids was ≤6.14 Å, which corresponded to the 95th percentile of the distribution of all pairwise distances between ligand centroids within each cluster in the PDB-LIG set (Fig. S31). We found that only 0.18% and 0.56% of Fpocket and P2rank predicted pocket pairs, respectively, had a distance between their centroids below that value. Given the apparent over-prediction of pockets of the two methods, we explored the precision/recall balance when keeping only the top scoring predicted pockets. Overall, we found that the best strategy to detect real binding sites in ligand-free structures was the combination of Fpocket detection and Prank scoring. Using the mentioned distance cut-off (6.14 Å) and considering the top-2 best scored pockets for each domain, we were able to detect 72% of the real pockets while 47% of detected pockets were indeed real (Fig. 2b).

Thus, we first ran Fpocket on the *apo* PDB reference structure for each domain to identify potential druggable pockets, we then ranked them by means of Prank, and we finally kept the top-2 ranked pockets per domain. Overall, this accounted for a total of 14,413 predicted pockets in 7403 domains (1806 unique Pfam domains in 4643 proteins). We named this set of pockets PDB-PD.

We then used the same strategy and criteria as before to detect pockets onto the predicted AF2 domain structures (Fpocket and Prank combination filtering out those pockets having residues with pLDDT values < 70), annotating a total of 32,202 pockets in 19,211 domains (2409 unique Pfam domains in 10,314 proteins). We named this set of pockets AF2-PD.

For each pocket and structure, we calculated pocket volume and buriedness using the rDock CAVITY functionality[62] and BioPython[94], respectively.

Assessment of the effect of protein flexibility

To exhaustively assess the behavior of PocketVec descriptors upon protein flexibility and conformational changes, from the PDB-LIG set, we selected those pockets for which we had i) ≥ 10 experimental structures with a co-crystallized ligand (i.e. *holo*), ii) ≥ 10 experimental structures without a ligand (i.e. *apo*) and iii) and AF2 structure with pLDDT <70 for all residues at a distance <8 Å from the pocket centroid. To avoid biasing the results towards overrepresented pockets, we capped the number of selected PDB structures to 10 in both cases (*holo* and *apo*). Overall, we kept 43 pockets and generated PocketVec descriptors for 10 of their *holo* structures, 10 of their *apo* structures and the corresponding AF2 model, resulting in a total number of 903 PocketVec descriptors (43 pockets x 21 descriptors).

Additionally, we gathered MD data from ATLAS[104], a repository of standardized MD simulations for 1390 PDB chains including their calculated trajectories as well as several quantitative analyses. In PDB-LIG (PDB-PD), we found available MD trajectories for 7 (96) of their 1279 (7403) domains, encompassing 8 (184) pockets. For each domain (103) we randomly sampled 9 MD frames from ATLAS trajectories (3 random frames per replica, x3 replicas), using the corresponding TPR and XTC GROMACS files. Thus, for each pocket (192), we considered 10 individual structures: the original one from the PDB-LIG/PDB-PD set and 9 frames from the MD simulations. Then, for each pocket, we calculated all pairwise distances (45) among their PocketVec descriptors (10) and compared them with background distances against each of the 4 sets of PocketVec descriptors we already precompiled (i.e. PDB-LIG, PDB-PD, AF2-LIG, AF2-PD).

Assessment of the effect of metals binding proteins

To assess the ability of our precomputed PocketVec descriptors to identify pocket similarities among metal-binding proteins, such as histone deacetylases (HDACs) or matrix metalloproteinases (MMPs). We first collected all domains within the PDB-LIG and PDB-PD sets (7399) and kept only those having a metal atom (AU, MN, NA, CU, AG, CO, PB, RB, K, SR, FE, MG, CD, NI, HG, CA, PT, TI, ZN or CS) in at least 1 associated PDB chain (3568). For further analysis, we only considered those PDB-LIG and PDB-PD pockets that were included within one of these domains (1052 out of 1594 and 6882 out of 14,368, respectively). After that and, for each pocket within each set, we individually considered whether any of the associated PDB chains had a metal atom at a distance <6.14 Å from the pocket centroid. For those that indeed had at least 1 metal atom (244 and 785 pockets for the PDB-LIG and PDB-PD sets, respectively) we kept it (if it was on the domain reference structure) or superimposed it (randomly selected from associated PDB chains) against the domain reference structure, and generated PocketVec descriptors for them with the explicit presence of the metal atom.

Systematic comparison of druggable pockets within domain families

First, for each pair of pockets within the AF2-PD and AF2-LIG sets that were located at the same Pfam domains, we computed the correlation between PocketVec similarity (defined as 1-PocketVec distance), sequence identity and Cα RMSD among pockets. We selected a maximum of 10 protein domains per Pfam domain to avoid biasing the results towards the most frequently occurring ones. We then removed those domain pairs having a global TM-score <0.5 and we performed all pairwise residue mappings using the corresponding Pfam multiple sequence alignments (MSA). After that, pockets (residues <8 Å) were compared on the basis of their sequences and structures (sequence identity and Cα RMSD, respectively). In fact, we only considered those pocket pairs having a sequence alignment coverage ≥80% and a centroid distance <6.14 Å after domain structural alignment, to account for structural variability (see *Pocket detection* in the previous section).

Additionally, we also ran an all-against-all comparison of pockets in the human pocketome (e.g. PDB-LIG, PDB-PD, AF2-LIG and AF2-PD). Global domain structures were compared using TM-align[103] (Cα RMSD, TM-score), while domain sequence identities were calculated using global pairwise alignments in BioPython[94] (Needleman-Wunsch algorithm, BLOSUM62, gap opening = −10, gap extension = −0.5).

Relationship between PocketVec similarity and experimentally determined compound-target pairs

We obtained raw binding data from the Chemical Checker compound-target space[64], which contains experimental data on 836,654 compounds bound to 6933 protein targets, as collected from ChEMBL v.33[75] and BindingDB v.2024[76]. We also collected all PocketVec descriptors from our study (i.e. PDB-LIG, PDB-PD; AF2-LIG and AF2-PD), encompassing a total number of 10,539 proteins. Of these, we had binding data for 2028 proteins. We then evaluated the corresponding 2,055,378 protein pairs on the bases of the number of shared compounds (in ChEMBL and BindingDB) and the cosine distances between their PocketVec descriptors.

We then calculated the number of protein pairs within each range of PocketVec distances in an all-vs-all comparison of PocketVec descriptors where, for each protein pair, we kept the minimal distance between PocketVec descriptors, regardless of their set of origin (e.g. PDB-LIG, AF2-PD, etc). Additionally, we depicted the number of shared compounds between each protein pair plotted against the minimal PocketVec distance between each pair and, finally, we calculated the Fisher test odds ratio considering a varying number of shared compounds for a PocketVec distance ≤ 0.17, ≤ 0.10 and ≤ 0.05.

We also performed a similar analysis on the fragment-protein binding pairs recently presented in ref. 105, where Winter and coworkers comprehensively tested the potential binding of 407 fragment compounds on 5951 proteins. We applied the filtering criteria defined by the authors in the original m/s (i.e. log2 fold change > 2.3, median normalized log2 fold change > 1, p-value < 0.05 and adjusted p-value < 0.25) and we ended up with a binary interaction matrix comprising 332 fragment compounds and 2744 proteins. As also indicated by the authors, we further removed proteins that were frequent hitters (>40 active fragments) as well as those that were rarely hit (<4 active fragments), restricting our analyses to 301 fragment compounds and 525 proteins for which we had PocketVec descriptors.

Comparison of kinase inhibition profiles with PocketVec descriptors, sequence and structure similarity measurements

We collected experimentally determined binding affinities from Klaeger et al.[82] and Reinecke et al.[83] (Supplementary Data 1). We considered undefined measures as inactive, and low-confidence and high-confidence measures were binarized at 30 nM (as recommended in Pharos[106]). We then removed all compounds that did not inhibit any kinase and all protein kinases that did not have any inhibitor or any PocketVec descriptor in the Protein Kinase Domain (Pfam PF00069). In this way, we eventually defined a binary inhibition matrix between 111 protein kinases and 94 small molecule kinase inhibitors for Klaeger et al.[82] (Fig. 7a and Supplementary Data 1a) and between 73 kinases and 164 inhibitors for Reinecke et al.[83] (Fig. 7b and Supplementary Data 1b).

We pairwise compared protein kinases on the basis of their binarized inhibition profiles (Jaccard similarity) and their PocketVec descriptors (employing the minimum distance among all PocketVec descriptors within their Protein Kinase Domains PF00069). Additionally, we also performed sequential and structural comparisons between kinases at domain level following the same strategy as in previous sections. In brief, domain structures were compared using TM-align[103] (Cα RMSD, TM-score) and domain sequence identities were calculated using global pairwise alignments in BioPython[94] (Needleman-Wunsch algorithm, BLOSUM62, gap opening = −10, gap extension = −0.5). Only the highest TM-score and sequence identity values among domains were considered for each pair of kinases.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data for all quantitative figures and depictions can be found in the form of TSV files within the Supplementary Data and Source Data files. All precomputed PocketVec descriptors and domain comparisons with sequence and structure similarity measurements are also provided in our Gitlab repository (https://gitlabsbnb.irbbarcelona.org/acomajuncosa/pocketvec). Finally, all PDB files used in this study for the structural characterization of human protein domains are exhaustively listed in Supplementary Data 2. Source data are provided with this paper.

Code availability

All the code necessary to generate PocketVec descriptors for any pocket of interest is available in our GitLab (https://gitlabsbnb.irbbarcelona.org/acomajuncosa/pocketvec), together with installation instructions and guidelines for running the code. On average, the generation of a PocketVec descriptor for a protein pocket takes 1 h on an Intel Xeon Gold 6138 CPU.

References

1. Sadybekov, A. V. & Katritch, V. Computational approaches streamlining drug discovery. *Nature* **616**, 673–685 (2023).
2. Goodsell, D. S. et al. RCSB Protein Data Bank: Enabling biomedical research and drug discovery. *Protein Sci.* **29**, 52–65 (2020).
3. Batool, M., Ahmad, B. & Choi, S. A Structure-Based Drug Discovery Paradigm. *IJMS* **20**, 2783 (2019).
4. Śledź, P. & Caflisch, A. Protein structure-based drug design: from docking to molecular dynamics. *Curr. Opin. Struct. Biol.* **48**, 93–102 (2018).
5. Macalino, S. J. Y., Gosu, V., Hong, S. & Choi, S. Role of computer-aided drug design in modern drug discovery. *Arch. Pharm. Res.* **38**, 1686–1701 (2015).
6. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).

7. Westermaier, Y., Barril, X. & Scapozza, L. Virtual screening: an in silico tool for interlacing the chemical universe with the proteome. *Methods* **71**, 44–57 (2015).
8. Lee, A., Lee, K. & Kim, D. Using reverse docking for target identification and its applications for drug discovery. *Expert Opin. Drug Discov.* **11**, 707–715 (2016).
9. Pinzi, L. & Rastelli, G. Molecular Docking: Shifting Paradigms in Drug Discovery. *IJMS* **20**, 4331 (2019).
10. Li, J., Fu, A. & Zhang, L. An Overview of Scoring Functions Used for Protein–Ligand Interactions in Molecular Docking. *Interdiscip. Sci. Comput Life Sci.* **11**, 320–328 (2019).
11. Shen, C. et al. From machine learning to deep learning: Advances in scoring functions for protein–ligand docking. *WIREs Comput Mol Sci* **10**, https://doi.org/10.1002/wcms.1429 (2020).
12. Sydow, D. et al. Advances and Challenges in Computational Target Prediction. *J. Chem. Inf. Model.* **59**, 1728–1742 (2019).
13. Fernández-Torras, A., Comajuncosa-Creus, A., Duran-Frigola, M. & Aloy, P. Connecting chemistry and biology through molecular descriptors. *Curr. Opin. Chem. Biol.* **66**, 102090 (2022).
14. Cereto-Massagué, A. et al. Molecular fingerprint similarity search in virtual screening. *Methods* **71**, 58–63 (2015).
15. Muegge, I. & Mukherjee, P. An overview of molecular fingerprint similarity search in virtual screening. *Expert Opin. Drug Discov.* **11**, 137–148 (2016).
16. Bileschi, M. L. et al. Using deep learning to annotate the protein universe. *Nat. Biotechnol.* **40**, 932–937 (2022).
17. Eguida, M. & Rognan, D. Estimating the Similarity between Protein Pockets. *Int. J. Mol. Sci.* **23**, https://doi.org/10.3390/ijms232012462 (2022).
18. Weill, N. & Rognan, D. Alignment-Free Ultra-High-Throughput Comparison of Druggable Protein–Ligand Binding Sites. *J. Chem. Inf. Model.* **50**, 123–135 (2010).
19. Schalon, C., Surgand, J.-S., Kellenberger, E. & Rognan, D. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins* **71**, 1755–1778 (2008).
20. Gainza, P. et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
21. Wood, D. J., de Vlieg, J., Wagener, M. & Ritschel, T. Pharmacophore Fingerprint-Based Approach to Binding Site Subpocket Similarity and Its Application to Bioisostere Replacement. *J. Chem. Inf. Model.* **52**, 2031–2043 (2012).
22. Desaphy, J., Raimbaud, E., Ducrot, P. & Rognan, D. Encoding Protein–Ligand Interaction Patterns in Fingerprints and Graphs. *J. Chem. Inf. Model.* **53**, 623–637 (2013).
23. Siragusa, L., Cross, S., Baroni, M., Goracci, L. & Cruciani, G. BioGPS: Navigating biological space to predict polypharmacology, off-targeting, and selectivity: Identifying Structurally Similar Sites through MIFs. *Proteins* **83**, 517–532 (2015).
24. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M. & Blaschke, T. The rise of deep learning in drug discovery. *Drug Discov. Today* **23**, 1241–1250 (2018).
25. Simonovsky, M. & Meyers, J. DeeplyTough: Learning Structural Comparison of Protein Binding Sites. *J. Chem. Inf. Model.* **60**, 2356–2366 (2020).
26. Scott, O. B., Gu, J. & Chan, A. W. E. Classification of Protein-Binding Sites Using a Spherical Convolutional Neural Network. *J. Chem. Inf. Model.* **62**, 5383–5396 (2022).
27. Morphy, R., Kay, C. & Rankovic, Z. From magic bullets to designed multiple ligands. *Drug Discov. Today* **9**, 641–651 (2004).
28. Konc, J. Binding site comparisons for target-centered drug discovery. *Expert Opin. Drug Discov.* **14**, 445–454 (2019).
29. Naderi, M. et al. Binding site matching in rational drug design: algorithms and applications. *Brief. Bioinforma.* **20**, 2167–2184 (2019).
30. Zhang, W., Pei, J. & Lai, L. Computational Multitarget Drug Design. *J. Chem. Inf. Model.* **57**, 403–412 (2017).
31. Haupt, V. J., Daminelli, S. & Schroeder, M. Drug Promiscuity in PDB: Protein Binding Site Similarity Is Key. *PLoS ONE* **8**, e65894 (2013).
32. Konc, J. & Janežič, D. Binding site comparison for function prediction and pharmaceutical discovery. *Curr. Opin. Struct. Biol.* **25**, 34–39 (2014).
33. Duran-Frigola, M. et al. Detecting similar binding pockets to enable systems polypharmacology. *PLoS Comput Biol.* **13**, e1005522 (2017).
34. Ehrt, C., Brinkjost, T. & Koch, O. Impact of Binding Site Comparisons on Medicinal Chemistry and Rational Molecular Design. *J. Med. Chem.* **59**, 4121–4151 (2016).
35. Jalencas, X. & Mestres, J. Identification of Similar Binding Sites to Detect Distant Polypharmacology. *Mol. Inf.* **32**, 976–990 (2013).
36. Salentin, S., Haupt, V. J., Daminelli, S. & Schroeder, M. Polypharmacology rescored: Protein–ligand interaction profiles for remote binding site similarity assessment. *Prog. Biophysics Mol. Biol.* **116**, 174–186 (2014).
37. Zhao, Z., Xie, L., Xie, L. & Bourne, P. E. Delineation of Polypharmacology across the Human Structural Kinome Using a Functional Site Interaction Fingerprint Approach. *J. Med. Chem.* **59**, 4326–4341 (2016).
38. Schumann, M. & Armen, R. S. Identification of Distant Drug Off-Targets by Direct Superposition of Binding Pocket Surfaces. *PLoS ONE* **8**, e83533 (2013).
39. Bongers, B. J., Ijzerman, A. P. & Van Westen, G. J. P. Proteochemometrics – recent developments in bioactivity and selectivity modeling. *Drug Discov. Today.: Technol.* **32-33**, 89–98 (2019).
40. D'Souza, S., Prema, K. V. & Balaji, S. Machine learning models for drug–target interactions: current knowledge and future directions. *Drug Discov. Today* **25**, 748–756 (2020).
41. Ehrt, C., Brinkjost, T. & Koch, O. A benchmark driven guide to binding site comparison: An exhaustive evaluation using tailor-made data sets (ProSPECCTs). *PLoS Comput Biol.* **14**, e1006483 (2018).
42. Jiménez-Luna, J., Grisoni, F., Weskamp, N. & Schneider, G. Artificial intelligence in drug discovery: recent advances and future perspectives. *Expert Opin. Drug Discov.* **16**, 949–959 (2021).
43. Vamathevan, J. et al. Applications of machine learning in drug discovery and development. *Nat. Rev. Drug Discov.* **18**, 463–477 (2019).
44. Ching, T. et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R Soc. Interf.* **15**, https://doi.org/10.1098/rsif.2017.0387 (2018).
45. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
46. Tunyasuvunakool, K. et al. Highly accurate protein structure prediction for the human proteome. *Nature* **596**, 590–596 (2021).
47. Varadi, M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**, D439–D444 (2022).
48. David, A., Islam, S., Tankhilevich, E. & Sternberg, M. J. E. The AlphaFold Database of Protein Structures: A Biologist's Guide. *J. Mol. Biol.* **434**, 167336 (2022).
49. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
50. Lipinski, C. & Hopkins, A. Navigating chemical space for biology and medicine. *Nature* **432**, 855–861 (2004).
51. Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discov. Today* **11**, 1046–1053 (2006).
52. Capecchi, A., Probst, D. & Reymond, J. L. One molecular fingerprint to rule them all: drugs, biomolecules, and the metabolome. *J. Cheminform* **12**, 43 (2020).

53. Govindaraj, R. G. & Brylinski, M. Comparative assessment of strategies to identify similar ligand-binding pockets in proteins. *BMC Bioinforma.* **19**, 91 (2018).
54. Schmidt, D. et al. Analyzing Kinase Similarity in Small Molecule and Protein Structural Space to Explore the Limits of Multi-Target Screening. *Molecules* **26**, https://doi.org/10.3390/molecules26030629 (2021).
55. Pu, L., Govindaraj, R. G., Lemoine, J. M., Wu, H.-C. & Brylinski, M. DeepDrug3D: Classification of ligand-binding pockets in proteins with a convolutional neural network. *PLoS Comput. Biol.* **15**, e1006718 (2019).
56. Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharm.* **152**, 5–7 (2007).
57. Keiser, M. J. et al. Relating protein pharmacology by ligand chemistry. *Nat. Biotechnol.* **25**, 197–206 (2007).
58. Falaguera, M. J. & Mestres, J. Illuminating the Chemical Space of Untargeted Proteins. *J. Chem. Inf. Model* **63**, 2689–2698 (2023).
59. Gao, M. & Skolnick, J. A Comprehensive Survey of Small-Molecule Binding Pockets in Proteins. *PLoS Comput Biol.* **9**, e1003302 (2013).
60. Friesner, R. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 1. Method and Assessment of Docking Accuracy. *J. Med. Chem.* **47**, 1739–1749 (2004).
61. Halgren, T. A. et al. Glide: A New Approach for Rapid, Accurate Docking and Scoring. 2. Enrichment Factors in Database Screening. *J. Med. Chem.* **47**, 1750–1759 (2004).
62. Ruiz-Carmona, S. et al. rDock: A Fast, Versatile and Open Source Program for Docking Ligands to Proteins and Nucleic Acids. *PLoS Comput Biol.* **10**, e1003571 (2014).
63. Koes, D. R., Baumgartner, M. P. & Camacho, C. J. Lessons Learned in Empirical Scoring with smina from the CSAR 2011 Benchmarking Exercise. *J. Chem. Inf. Model.* **53**, 1893–1904 (2013).
64. Duran-Frigola, M. et al. Extending the small-molecule similarity principle to all levels of biology with the Chemical Checker. *Nat. Biotechnol.* **38**, 1087–1096 (2020).
65. Bertoni, M. et al. Bioactivity descriptors for uncharacterized chemical compounds. *Nat. Commun.* **12**, 3932 (2021).
66. Fernandez-Torras, A., Duran-Frigola, M., Bertoni, M., Locatelli, M. & Aloy, P. Integrating and formatting biomedical data as pre-calculated knowledge graph embeddings in the Bioteque. *Nat. Commun.* **13**, 5304 (2022).
67. The UniProt, C. et al. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
68. Mistry, J. et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412–D419 (2021).
69. Zhang, Y. et al. Benchmarking Refined and Unrefined AlphaFold2 Structures for Hit Discovery. *J. Chem. Inf. Model* **63**, 1656–1667 (2023).
70. Holcomb, M., Chang, Y. T., Goodsell, D. S. & Forli, S. Evaluation of AlphaFold2 structures as docking targets. *Protein Sci.* **32**, e4530 (2023).
71. Scardino, V., Di Filippo, J. I. & Cavasotto, C. N. How good are AlphaFold models for docking-based virtual screening? *iScience* **26**, 105920 (2023).
72. Lyu, J. et al. AlphaFold2 structures template ligand discovery. *bioRxiv*, https://doi.org/10.1101/2023.12.20.572662 (2023).
73. Le Guilloux, V., Schmidtke, P. & Tuffery, P. Fpocket: an open source platform for ligand pocket detection. *BMC Bioinforma.* **10**, 168 (2009).
74. Krivak, R. & Hoksza, D. P2Rank: machine learning based tool for rapid and accurate prediction of ligand binding sites from protein structure. *J. Cheminform* **10**, 39 (2018).
75. Zdrazil, B. et al. The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* **52**, D1180–D1192 (2024).
76. Gilson, M. K. et al. BindingDB in 2015: A public database for medicinal chemistry, computational chemistry and systems pharmacology. *Nucleic Acids Res.* **44**, D1045–D1053 (2016).
77. Attwood, M. M., Fabbro, D., Sokolov, A. V., Knapp, S. & Schioth, H. B. Trends in kinase drug discovery: targets, indications and inhibitor design. *Nat. Rev. Drug Discov.* **20**, 839–861 (2021).
78. Roskoski, R. Jr. Properties of FDA-approved small molecule protein kinase inhibitors: A 2023 update. *Pharm. Res.* **187**, 106552 (2023).
79. Karaman, M. W. et al. A quantitative analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **26**, 127–132 (2008).
80. Davis, M. I. et al. Comprehensive analysis of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1046–1051 (2011).
81. Hanson, S. M. et al. What Makes a Kinase Promiscuous for Inhibitors? *Cell Chem. Biol.* **26**, 390–399.e395 (2019).
82. Klaeger, S. et al. The target landscape of clinical kinase drugs. *Science* **358**, https://doi.org/10.1126/science.aan4368 (2017).
83. Reinecke, M. et al. Chemical proteomics reveals the target landscape of 1,000 kinase inhibitors. *Nat. Chem. Biol.* https://doi.org/10.1038/s41589-023-01459-3 (2023).
84. Wang, S. et al. CavitySpace: A Database of Potential Ligand Binding Sites in the Human Proteome. *Biomolecules* **12**, https://doi.org/10.3390/biom12070967 (2022).
85. Konc, J. & Janezic, D. ProBiS-Fold Approach for Annotation of Human Structures from the AlphaFold Database with No Corresponding Structure in the PDB to Discover New Druggable Binding Sites. *J. Chem. Inf. Model* **62**, 5821–5829 (2022).
86. Sim, J., Kwon, S. & Seok, C. HProteome-BSite: predicted binding sites and ligands in human 3D proteome. *Nucleic Acids Res.* **51**, D403–D408 (2023).
87. Tsuchiya, Y. et al. PoSSuM v.3: A Major Expansion of the PoSSuM Database for Finding Similar Binding Sites of Proteins. *J. Chem. Inf. Model* **63**, 7578–7587 (2023).
88. Carter, A. J. et al. Target 2035: probing the human proteome. *Drug Discov. Today* **24**, 2111–2115 (2019).
89. Jin, W., Barzilay, R. & Jaakkola, T. S. Hierarchical Generation of Molecular Graphs using Structural Motifs. *arXiv*, (2020).
90. Blaschke, T. et al. REINVENT 2.0: An AI Tool for De Novo Drug Design. *J. Chem. Inf. Model* **60**, 5918–5922 (2020).
91. 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.* **27**, 675–679 (2013).
92. Wong, F. et al. Discovery of a structural class of antibiotics with explainable deep learning. *Nature*, https://doi.org/10.1038/s41586-023-06887-8 (2023).
93. Jayatunga, M. K. P., Xie, W., Ruder, L., Schulze, U. & Meier, C. AI in small-molecule drug discovery: a coming wave? *Nat. Rev. Drug Discov.* **21**, 175–176 (2022).
94. Cock, P. J. A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
95. Andrio, P. et al. BioExcel Building Blocks, a software library for interoperable biomolecular simulation workflows. *Sci. Data* **6**, 169 (2019).
96. Word, J. M., Lovell, S. C., Richardson, J. S. & Richardson, D. C. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J. Mol. Biol.* **285**, 1735–1747 (1999).
97. O'Boyle, N. M. et al. Open Babel: An open chemical toolbox. *J. Cheminform* **3**, 33 (2011).
98. Harris et al. Array programming with NumPy. *Nature* **585**, 357–362 (2020).
99. Virtanen, P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* **17**, 261–272 (2020).

100. Kahraman, A., Morris, R. J., Laskowski, R. A. & Thornton, J. M. Shape Variation in Protein Binding Pockets and their Ligands. *J. Mol. Biol.* **368**, 283–301 (2007).
101. Barelier, S., Sterling, T., O'Meara, M. J. & Shoichet, B. K. The Recognition of Identical Ligands by Unrelated Proteins. *ACS Chem. Biol.* **10**, 2772–2784 (2015).
102. Ludwiczak, J., Winski, A. & Dunin-Horkawicz, S. localpdb-a Python package to manage protein structures and their annotations. *Bioinformatics* **38**, 2633–2635 (2022).
103. Zhang, Y. & Skolnick, J. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res.* **33**, 2302–2309 (2005).
104. Vander Meersche, Y., Cretin, G., Gheeraert, A., Gelly, J. C. & Galochkina, T. ATLAS: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Res* **52**, D384–D392 (2024).
105. Offensperger, F. et al. Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells. *Science* **384**, eadk5864 (2024).
106. Nguyen, D. T. et al. Pharos: Collating protein information to shed light on the druggable genome. *Nucleic Acids Res.* **45**, D995–D1002 (2017).

Acknowledgements

P.A. acknowledges the support of the Generalitat de Catalunya (RIS3-CAT Emergents CECH: 001-P-001682 and VEIS: 001-P-001647; and 2021 SGR 00876), the Spanish Ministerio de Ciencia, Innovación y Universidades (PID2020-119535RB-I00), the Instituto de Salud Carlos III (IMPaCT-Data), and the European Commission (RiPCoN: 101003633). A.C-C. is a recipient of an FI fellowship (2020 FI_B 00094). We also acknowledge institutional funding from the Spanish Ministry of Science and Innovation through the Centres of Excellence Severo Ochoa Award, and from the CERCA Programme / Generalitat de Catalunya.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-52146-3.

Correspondence and requests for materials should be addressed to Patrick Aloy.

Peer review information *Nature Communications* thanks Shovanlal Gayen, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.