Excuse me, there is a mutant in my bioactivity soup!

A comprehensive analysis of the genetic variability landscape of bioactivity databases and its effect on activity modelling

Marina Gorostiola González[1,2,‡], Olivier J. M. Béquignon[1,‡], Emma J. Manners[3], Anna Gaulton[3], Prudence Mutowo[3], Elisabeth Dawson[3], Barbara Zdrazil[3], Andrew R. Leach[3], Adriaan P. IJzerman[1], Laura H. Heitman[1,2], and Gerard J. P. van Westen*[,1]

[1]*Leiden Academic Centre of Drug Research, Leiden University, 55 Einsteinweg, 2333 CC. Leiden, The Netherlands*
[2]*Oncode Institute, The Netherlands*
[3]*European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SD, United Kingdom*
[‡]*These authors contributed equally*
*Corresponding author (gerard@lacdr.leidenuniv.nl)

Abstract

Bioactivity prediction is essential in computational drug discovery, particularly within virtual screening campaigns. Despite advancements in model architectures and features, the sparsity and quality of relevant training data remain a major bottleneck. Notably, genetic variance annotation, crucial for understanding variant-specific bioactivity, is often neglected. Key efforts to tackle these issues are conducted by public bioactivity databases such as ChEMBL, but these are not free of challenges. Here, a comprehensive analysis of the extent and distribution of bioactivity data tested on genetic variants across organisms, protein families, individual targets, and specific variants, for the first time characterises in detail the genetic variability landscape in the ChEMBL database and sheds light on the range and consequences of protein amino acid substitutions in bioactivity data distribution and modelling. Furthermore, an extensive set of analysis resources (Python package and notebooks) and a variant-annotated bioactivity dataset are made available to help replicate the analyses described here for any protein of interest and make informed decisions regarding the quality of data for modelling. Finally, the potential to extract variants and subsets of the chemical space with desirable inter-variant bioactivity profiles is demonstrated for data-rich proteins. This approach contributes to more reliable bioactivity modelling, aids noise reduction and informs decision-making in computational drug discovery.

Keywords: genetic variants, mutants, ChEMBL, activity data, modelling, QSAR, PCM

Introduction

Bioactivity prediction is one of the key techniques in the computational drug discovery pipeline, mostly applied in virtual screening campaigns[1,2]. Quantitative structure-activity relationship (QSAR) modelling has been around for a long time and can be used to predict ligand bioactivity for a target of interest based on the compound's chemical structural characteristics[3]. Over time other bioactivity prediction strategies have emerged that include information other than chemistry-derived features[4–8]. An example is proteochemometric (PCM) modelling, where the protein characteristics are considered in addition to ligand molecular structure, allowing for bioactivity predictions on several targets simultaneously[8–10].

Every year an increasing number of articles showcase improvements in machine learning and artificial intelligence (AI/ML) bioactivity modelling in the form of novel model architectures or chemical and protein descriptors, among other innovations.[11–16] Still, previous literature shows that one of the main bottlenecks in bioactivity prediction is the amount and quality of the available data for model training and testing[17,18]. Several databases, such as ChEMBL and PubChem, aim to compile as much data as possible by extracting it from the literature or accepting deposited datasets, which on its own can introduce errors[19,20]. Certain annotations like assay cell type, tissue, or genetic variants are not present in all articles or are described differently. In turn, this can result in inconsistencies in information content that affect the quality and comprehensiveness of the data[21,22].

Variant annotation in particular is one of the key aspects that should be considered when analysing bioactivity data[23]. The same compound can have a very different bioactivity on different genetic variants of the same protein[24–27]. In fact, some compounds are explicitly designed to have differential bioactivity across variants to, for example, reduce side effects by avoiding targeting the wild-type (WT) protein in anti-cancer therapies, or to target escape variants in antibiotics or antivirals[28,29]. However, variant annotation tends to be overlooked in bioactivity databases where, in many cases, it is not present or lacks validation. Moreover, even when variants are annotated - as is the case in the ChEMBL database - they are often ignored when constructing a bioactivity dataset, which only recently has been explicitly described as a potential source of noise[30,31]. The advantage of modelling variant-annotated data has been demonstrated in variant-rich organisms, such as HIV[32], and the implications in human proteins could be similarly important.

Here, we thoroughly evaluate the risks and opportunities presented by variant annotation in bioactivity databases by extensively characterising variant-annotated bioactivity data in the ChEMBL 31 database. Through an assessment of annotation fidelity, the non-triviality of this task is highlighted, and adjustments are proposed to improve the ChEMBL variant annotation

pipeline for future releases. A revised bioactivity dataset with protein amino acid substitution annotations is derived from this work and enriched with curated data from literature[33] (Christmann subset) previously curated as part of the Papyrus dataset[34]. The additional data is aggregated in this work with the ChEMBL annotated data following the pipeline with rigorous data curation and filtering, and standardization of molecular structures that were applied to obtain the Papyrus dataset. Furthermore, we investigate the distribution of variant-annotated bioactivity data points in the combined dataset across organisms, protein families, individual targets, and specific variants; and evaluate the effect of variants in bioactivity distribution and modelling. These findings not only contribute to advancing our understanding of the effects of amino acid substitutions in bioactivity but also provide invaluable insights for refining bioactivity data curation practices, particularly concerning variants, for enhanced predictive modelling purposes. Our work also highlights the importance of reporting comprehensively the full sequences of proteins used in bioassays and bioactivity measurements, in both the literature and when depositing data directly into databases.

Results

Variant annotation in bioactivity databases is far from trivial

Genetic variants are currently annotated in the ChEMBL database by manually extracting this information from the original articles for data originating from the scientific literature. Since ChEMBL 22 this information has been mapped to protein targets (alongside their UniProt accessions) and made available in a structured format via the *variant_sequences* table. In this work, an orthogonal approach has been used to evaluate the fidelity and comprehensiveness of these annotations and to include as many variants as possible for the analysis of bioactivities against proteins carrying amino acid substitutions (**Figure 1,** steps 1-7). This approach is expert knowledge-agnostic and embodies an automatic pipeline based exclusively on data previously extracted from the database. Its first step consisted of the automatic extraction of amino acid substitution patterns from the assay descriptions of unique pairs of assays and protein targets, and their subsequent validation against the WT protein sequence (**Figure 1,** step 2). The extracted substitutions were then compared to the ChEMBL variant annotations in a feedback loop in which mismatches were semi-automatically classified and used to rescue or revert annotations (**Figure 1c,** step 3). Finally, variant targets were annotated based on this feedback and mapped to ChEMBL bioactivity data. The final variant-enhanced bioactivity dataset (VEBD) was constructed by keeping exclusively bioactivity data for proteins with at least one variant annotated and was lastly enriched with variant-annotated bioactivity data from the Christmann dataset.

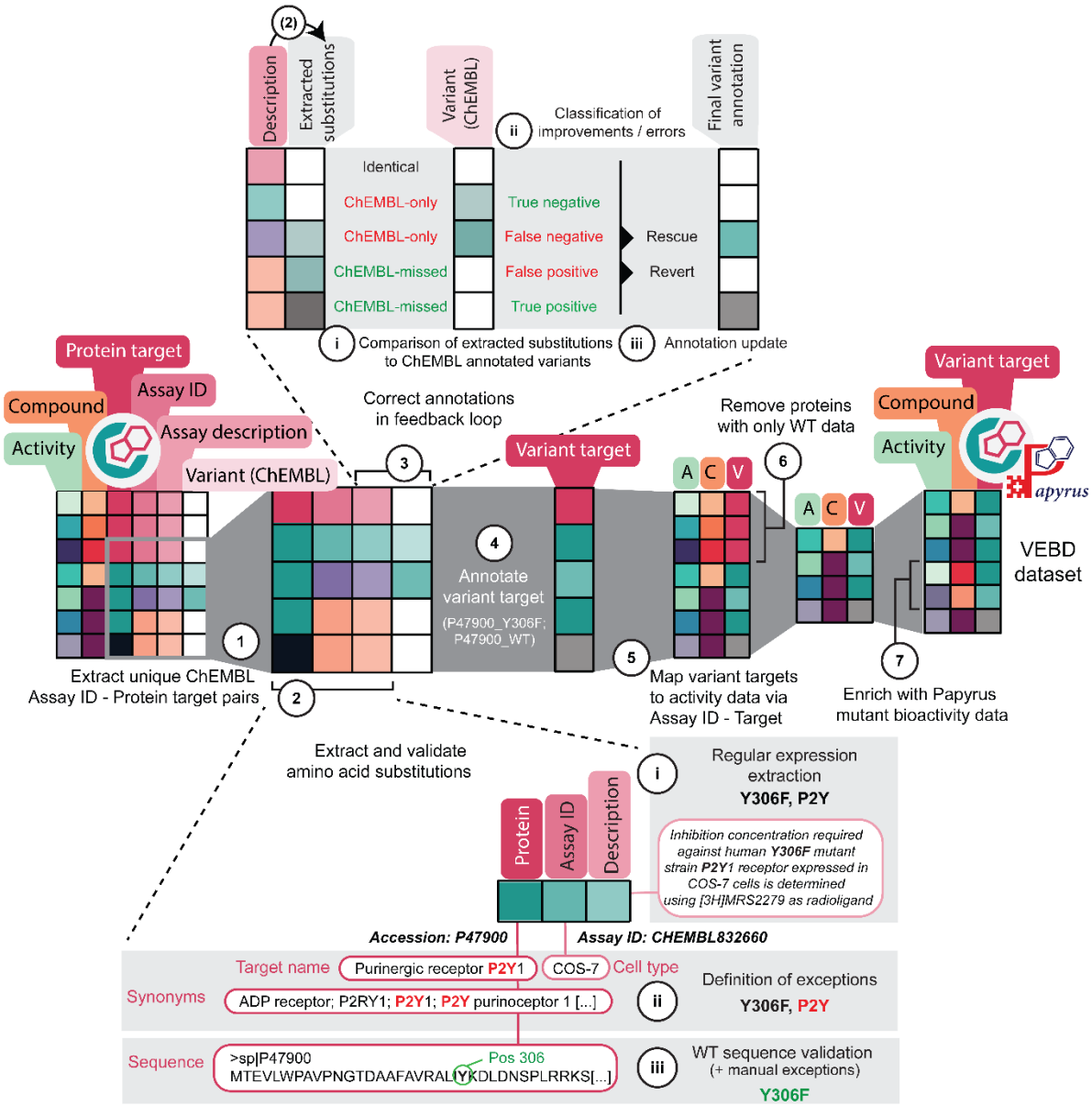

Figure 1. Pipeline to construct the variant-enhanced bioactivity dataset (VEBD) from ChEMBL and Papyrus data. (1) Unique assay-target pairs with bioactivity data are extracted from ChEMBL 31. (2) Regular expressions are used to extract amino acid substitution patterns, which are validated by introducing exceptions and mapping them to wild-type (WT) sequences. (3) Extracted substitutions are compared to ChEMBL annotated variants, and the classification of mismatches is used to determine the final annotations. More details of this step are available in Supplementary Figure 1. (4) A variant target identifier is defined based on the final variant annotations. (5) The variant target identifier is mapped back to the ChEMBL bioactivity dataset. (6) Proteins with only WT data are filtered out. (7) The bioactivity dataset is standardized and curated similarly to, and enriched with variant data from the Papyrus dataset.

Regular expressions were used to extract amino acid substitution patterns from assay descriptions, starting from 376,233 assay-protein target pairs in the ChEMBL 31 database with data suitable for regression modelling. Assay descriptions are extracted and curated in ChEMBL from the primary literature sources in a combined manual and semi-automated pipeline. Of note, genetic alterations other than amino acid substitutions were deemed out of

the scope for the initial stage of this project. As exemplified in **Figure 1** (step 2) for the assay-target pair CHEMBL832660 - P47900, these expression patterns could extract true substitutions, such as *Y306F,* but also incorrect patterns from the assay description, like *P2Y*. This first step yielded potential substitutions in 52,922 assay-target pairs. Therefore, exceptions were defined from other fields related to the assay and the target protein, in particular cell type, target preferred name, and target synonyms. This helped to refine the pipeline by rejecting extracted patterns such as *P2Y* that map to a part of the name of the assay target (purinoceptor P2Y1 in this case, UniProt accession P47900) and does not refer to a proline to tyrosine substitution. Indeed, 34,676 assay-target pair substitutions raised at least one exception flag. Of note, these exceptions are less of an issue in the original ChEMBL variant annotation pipeline, since some manual curation is performed. The substitution patterns that had not been flagged as exceptions were validated in the next step by checking the existence of the WT amino acid at the specified position in the target sequence. For example, in the case of the aforementioned *Y306F* substitution pattern, P2Y1 has indeed a tyrosine residue at position 306 of its sequence, hence this extracted substitution was validated. At this point, several additional exceptions were introduced by extracting patterns that were likely to be falsely validated, such as *M1*, as substitutions are unlikely to appear at the first position of the sequence, yet they would be given a false valid status as the starting codon AUG codes for methionine. This resulted in 8,455 assay-target pairs with WT sequence-validated extracted substitutions.

Next, the extracted and validated substitutions were compared to the originally annotated ChEMBL variants for all assay-target pairs (**Figure 1**, step 3**, Supplementary Figure 1**). This step, which we refer to as the annotation feedback loop, was included for three reasons, namely 1) to pinpoint highlights and pitfalls, 2) suggest improvements to the ChEMBL variant annotation pipeline, and 3) to include additional ChEMBL variants and collect the most complete dataset with variant annotated data in the scope of this project. Additionally, it served as a reminder of the non-triviality of the variant annotation process. Given its complexity, the feedback loop is now under review and remains subject to revision. The updated results will be incorporated in a revised version, therefore it is advisable to approach the following preliminary results with caution. Out of the 8,455 assay-target pairs with extracted substitutions, 7,622 (90%) had an identical annotation in ChEMBL. The remaining 833 were missing in ChEMBL, either completely (651) or because they had been flagged as “Undefined mutation” (182). Mismatching variants were further classified to determine their suitability for the VEBD (**Supplementary Table 1, Supplementary Figure 1**). Assays assessing more than one target were rejected for this analysis, as well as assays testing targets with variation corresponding to alterations or genotypes with ambiguous definitions. If a multiple substituted

protein was only partially validated; the annotations were rejected. If a validated amino acid substitution was combined with an insertion/deletion/truncation then the substitution was included for this analysis. Finally, non-substitution patterns that had been incorrectly validated against the WT sequence were identified as potential novel exceptions for improving the pipeline. Subsequently, these 833 entries were manually classified into 648 true positives that represent potential novel annotations missed by ChEMBL, and 185 false positives that arise from substitution extraction errors and will be used to refine the current pipeline. The true positive group was included in the final VEBD. Of note, among these were assay-target pairs with either completely novel extracted substitutions or rescues from previously undefined variants that were not fully annotated but were deemed inside the scope of this project. For example, we deemed as within scope, variants with co-occurring amino acid substitutions and deletions/duplications, flagged by ChEMBL as undefined variants and “rescued” for this project.

Apart from the 8,455 assay-target pairs with extracted substitutions, 1,600 pairs were found to be annotated only in ChEMBL and not identified by the current variant annotation pipeline. These ChEMBL-only annotated pairs were further evaluated in light of the underlying reasons that led to their exclusion from the current variant annotation pipeline (**Supplementary Table 2**, **Supplementary Figure 1**). ChEMBL substitutions missed by the regular expression, such as those with unconventional definitions, were incorporated into this analysis unless their initial annotation was "undefined" or a deletion. Extracted substitutions failing validation against the WT sequence were categorised into three groups: 1) If the extracted substitutions matched those in ChEMBL in all aspects except the residue number, the original substitutions were considered a sequence number shift exception and included. 2) If the extracted substitutions fully matched the original ChEMBL annotation but were not valid according to the WT sequence, they were either a) excluded (i.e. if the associated target was a protein family) or b) classified as ambiguous due to a sequence mismatch. 3) Finally, if the extracted substitutions did not align with the original annotation, they were deemed ambiguous due to substitution mismatch or omission and are under review. This analysis led to the classification of ChEMBL-only variants into true negatives (686 misclassified ChEMBL annotations), false negatives (798 ChEMBL expert annotations), and ambiguous (416 ChEMBL-only annotations). True negatives were excluded from the final dataset, while false negatives were rescued from ChEMBL and included. Pairs in the ambiguous group were flagged as undefined variants and included in the final dataset. After the annotation feedback loop, 9,229 assay-target pairs (774 additional assays) were annotated with variants. These were annotated with a variant target identifier as done in the Papyrus dataset by adding the amino acid substitutions as a suffix to the UniProt accession code of the protein. Similarly, bioactivity data points tested

on WT proteins were identified by the suffix “WT” after the accession code. Note that the final number of annotated pairs relies on the feedback loop, which is currently under revision; thus, the ultimate count is subject to change in an updated version.

To construct the VEBD, the variant target identifiers were mapped to ChEMBL bioactivity data based on assay-target pairs. Duplicated data from several assays for the same variant target were joined into one single point by dropping data with questioned validity, considered low-quality, and calculating the mean pChEMBL value or the most common activity flag. This resulted in 1,870,748 compound-target pairs across 6,777 targets, of which 25,259 contained variant targets - 736 with undefined variants - and the rest were WT. The ChEMBL 31 annotated set was merged with the fraction of the Papyrus dataset version 5.5 originating from the Christmann subset, keeping only targets with at least one variant defined for the follow-up analysis of variant-annotated bioactivity. The final combined VEBD for bioactivity analysis contained 455,839 compound-target pairs across 335 proteins, of which 25,086 data points represented data on variant proteins. Of these, 22,992 compound-target pairs originated from ChEMBL 31, 672 from the Papyrus Christmann subset, which were not present in ChEMBL, and 1,422 from both sources. In the following sections, we explore in detail the VEBD.

Variants are heterogeneously represented in bioactivity datasets across protein families

The first observation from review of the VEBD was that bioactivity data points were not homogeneously distributed across protein families. Proteins were assigned to their corresponding protein families using the levels L1-L5 in the protein family classification table in ChEMBL. Out of the 455,839 bioactivity data points in the VEBD, more than half were in enzymes (266,328), followed by membrane receptors (96,037) and then the remaining protein families (**Figure 2a, Supplementary Table 3**). The percentage of variant-tested bioactivity data with respect to the total amount of bioactivity data – hereby referred to as variant bioactivity percentage – was highest for secreted proteins (10.8%) followed by enzymes (7.8%), but in both cases, it was in the same order of magnitude as the variant bioactivity percentage for the whole dataset (5.5%). Of note, the secreted proteins family included only one protein while the enzymes family included 195. Compared to the highest classification level of protein families, the variant load drastically differed between protein subfamilies. For example, the variant bioactivity percentage across subfamilies of the kinase enzyme family ranged from 0.1% for the CMGC protein kinase group to 35% for the TKL protein kinase group (**Figure 2c, Supplementary Table 4**).

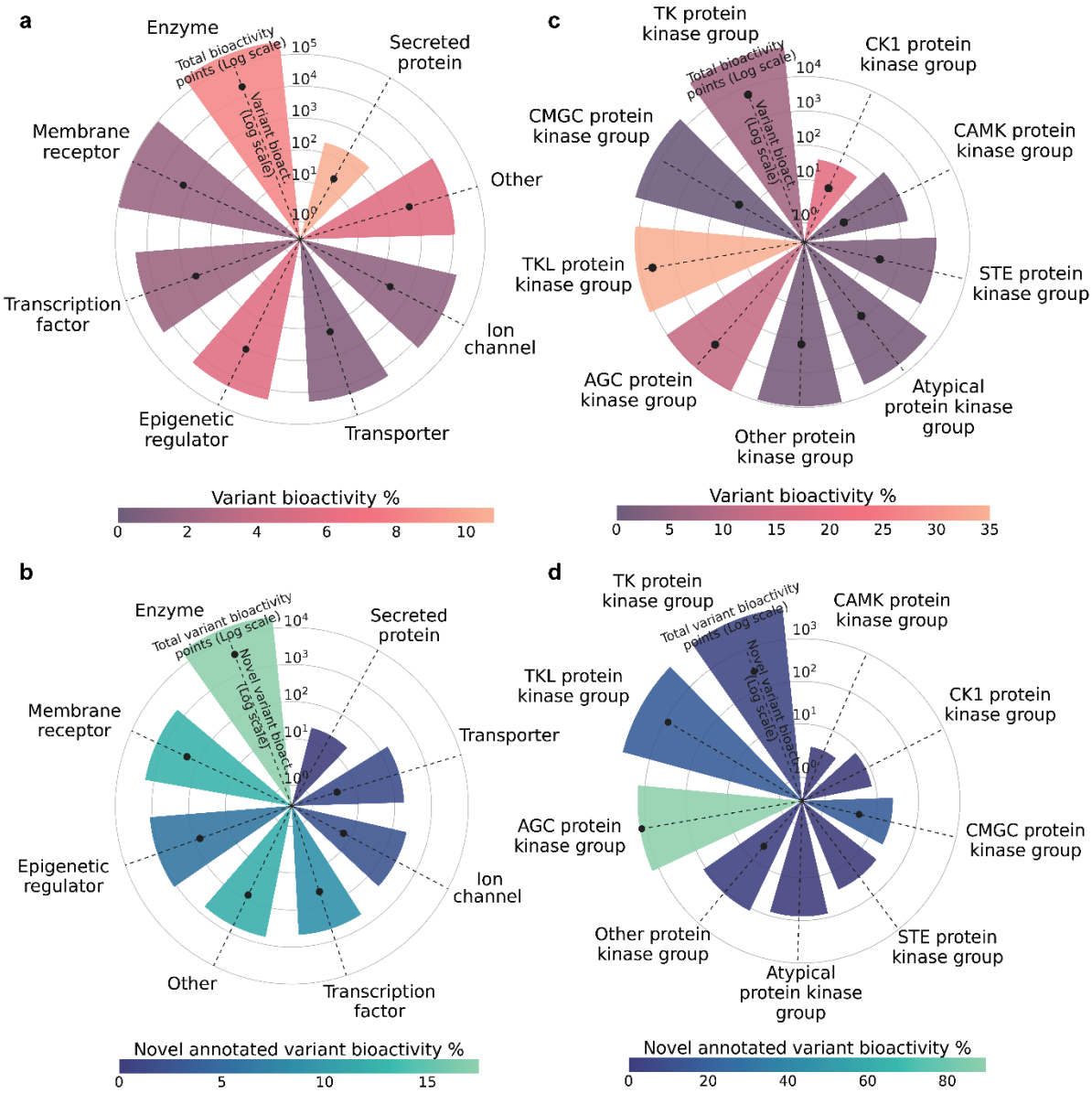

Figure 2. Distribution of variant bioactivity data across protein families in targets with at least one annotated variant. **a)** Bioactivity data in the VEBD for all protein families (L1 classification). **b)** Comparison of originally ChEMBL-annotated and novel variant data for all protein families (L1 classification). **c)** Bioactivity data in the VEBD for subfamilies of the Kinase enzymes family (L4 classification for L2 = Kinase). **d)** Comparison of originally ChEMBL-annotated and novel variant data for subfamilies of the Kinase enzymes family (L4 classification for L2 = Kinase). Bar heights represent the number of total bioactivity points (in a,c) or total variant bioactivity points (in b,d) on a logarithmic scale. The height of the black dots along the dashed lines represents the number of variant bioactivity points (in a,c) or novel annotated variant bioactivity points (in b,d). The colour gradient represents the percentage of variant bioactivity data with respect to total bioactivity data (in a,c) or the percentage of novel annotated variant data with respect to total variant bioactivity data (in b,d).

Similar trends were observed while focusing only on ChEMBL-exclusive data and exploring the differences between the original and the current variant annotation pipelines. The highest amount of bioactivity data points with potential novel variant annotations corresponded to enzymatic targets (3,631), followed by membrane receptors (218). However, at the highest

protein classification level, the percentage of potentially novel annotated bioactivity data to the totality of the variant-annotated data significantly differed across protein families, ranging from 0% in secreted proteins to 17.5% in enzymes (**Figure 2b, Supplementary Table 5**). Again, this effect was exacerbated across kinase subfamilies. Here, in four subfamilies (i.e. atypical, STE, CK1, and CAMK protein kinase groups) the totality of the variant bioactivity data had previously been annotated in ChEMBL, resulting in a novel annotated variant bioactivity percentage of 0%, while in the AGC protein kinase group, 89.7% of the variant data was introduced by the current variant annotation pipeline (**Figure 2d, Supplementary Table 6**). Similarly, in the kinase subfamily with the highest amount of variant data (i.e. TK protein kinase group), 5.1% of the variant data had not been previously annotated in ChEMBL.

The distribution of data in the VEBD across individual proteins was similarly unbalanced. Of the 335 proteins included in the annotated dataset, eight viral and bacterial proteins and one human protein did not include any WT data. However, only three of these (Hepatitis C viral NS3 protease Q0ZMF1 and polyprotein K7XJL6, and Human immunodeficiency virus 1 – HIV-1 – reverse transcriptase Q9WKE8) had more than 30 bioactivity data points. From the remaining 326 proteins, the vast majority (315) had simultaneously less than 20 variants and less than 10,000 bioactivity data (**Figure 3**, **Supplementary Table 7**). Only three human proteins (aldehyde dehydrogenase AL1A1 - P00352, phosphatidylinositol kinase PK3CA - P42336, and epidermal growth factor receptor EGFR - P00533) had more than 10,000 bioactivity data points, of which only one (EGFR) had a variant bioactivity percentage over 2%, specifically 18.36%. Moreover, eight different proteins had more than 20 annotated variants, including WT (**Figure 3a**). Some of these variants were single amino acid substitutions, while other variants accumulated several substitutions (**Supplementary Table 8**). The two most tested proteins among these eight with high genetic variance were viral proteins from HIV-1, namely polyprotein RNase H - reverse transcriptase (RNaseH-RT, Q72547) and polyprotein Q72874. The other six were mammalian membrane proteins, some of which may have been subjected to experimental mutagenesis programmes: five class A G protein-coupled receptors (GPCRs) – the human gonadotropin-releasing hormone receptor GNRHR (P30968), the rat muscarinic receptor ACM3 (P08483), the human chemokine receptor CXCR4 (P61073), the rat opioid receptor OPRK (P34975), and P2Y1 (P47900) – and one solute carrier transporter – human betaine transporter S6A12 (P48065). The protein with the largest number of annotated variants was GNRHR, with 70 variants other than the WT. Among the eight proteins with high genetic variance, the variant bioactivity percentages ranged between 1.72% and 71.83%.

From the 315 proteins that had simultaneously less than 20 variants and less than 10,000 bioactivity data, only 100 displayed a variant bioactivity percentage equal to or greater than 10% (**Figure 3b**), and only 10 of these had more than 1,000 bioactivity data points. For reference, we consider 1,000 data points as an arbitrary threshold to enable bioactivity modelling. Constraining the variant bioactivity percentage to 20% resulted in only 62 proteins out of which only six had more than 1,000 bioactivity data points; most of these contained clinically-relevant mutations. The five proteins with the largest amount of bioactivity data were all tyrosine, tyrosine-like, or AGC kinases, namely ABL1 (P00519), BRAF (P15056), leucine-rich repeat kinase LRRK2 (Q5S007), ALK (Q9UM73), and ribosomal protein kinase RPKS6B1 (P23443) in descending order of bioactivity data points and in line with the distributions per protein family (**Figure 2a,c)**. The sixth protein was the oxidoreductase isocitrate dehydrogenase IDHC (O75874).

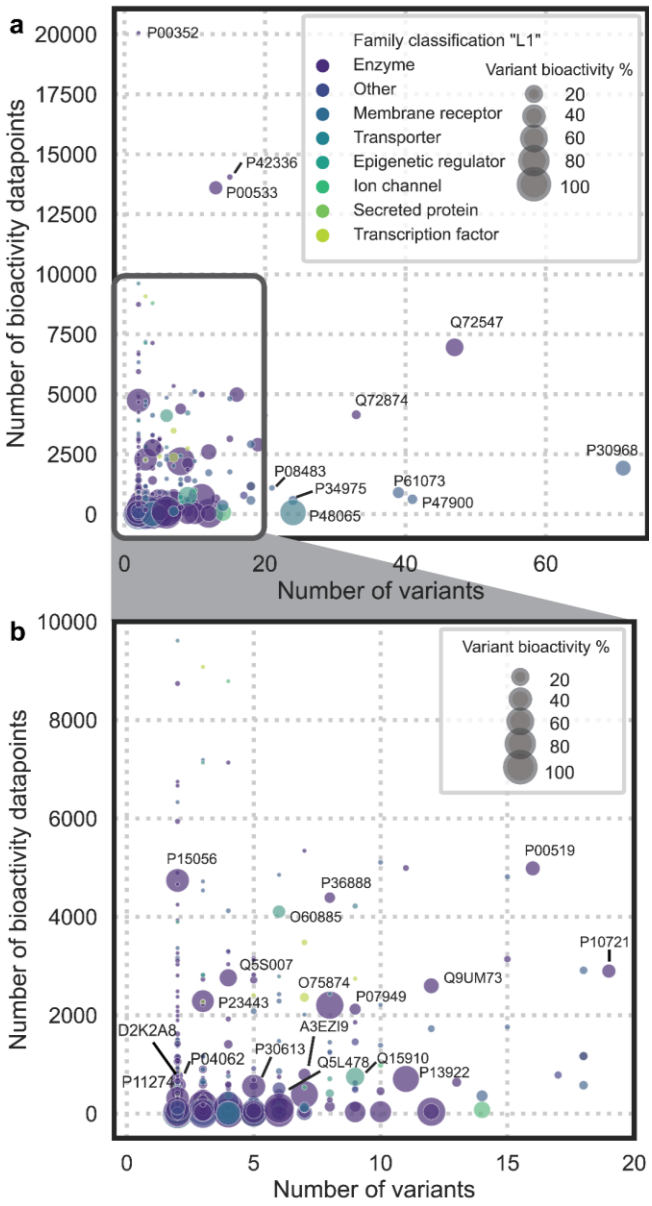

Figure 3. Variant annotation load per protein in terms of the number of variants and bioactivity data, as well as variant bioactivity percentage. **a)** Overall. Labelled are proteins with more than 10,000 data and/or more than 20 annotated variants, including WT. **b)** Proteins with less than 10,000 bioactivity data and less than 20 variants. Labelled are proteins with a variant bioactivity percentage higher than 10% and more than 500 data.

Save for the exceptions mentioned above, generally higher variant bioactivity percentages correlated with lower total bioactivity data, regardless of the number of variants annotated (**Supplementary Figure 2d**). From the total of 335 proteins in the dataset, only 32 showed as much or more bioactivity data for variants than for WT (i.e. 50% variant bioactivity percentage or higher), and out of these, only three had more than 1,000 bioactivity data points, namely IDHC (seven variants apart from WT), BRAF (one variant), and RPKS6B1 (two variants), and variant bioactivity percentages of 86.29%, 60.27%, and 55.21%.

In general, annotated proteins with more than 1,000 data points had a small number of variants, and most of their data was tested on the WT protein (**Supplementary Figure 2d**). However, the data-rich protein targets highlighted in this section emphasised the potential relevance of hidden variant data in bioactivity modelling and were therefore the focus for the rest of the analysis. In particular, we defined a set of 13 data-rich proteins (**Table 1**) with the highest variant bioactivity percentages (i.e. equal or above 10%) that had simultaneously sufficient data for bioactivity modelling (i.e. equal or above 1,000 bioactivity data points) and that were subsequently analysed in more detail in the following sections.

Amino acid substitution types represented in bioactivity datasets align with organism mutation rates

The type of amino acid substitutions represented in bioactivity datasets was also not homogeneously represented and may reflect the community's interest in protein variant sampling. As such, the majority of the reported variants were amino acid substitutions to alanine (**Figure 4a**), as part of the commonplace alanine scanning strategies to determine key structural and functional residues. Indeed, as expected, the alanine enrichment was not maintained in the number of bioactivity data points (**Figure 4b**). Instead, biologically relevant variants such as cancer-related BRAF V600E and EGFR T790M and L858R were responsible for the largest density of bioactivity data around particular amino acid substitutions. For example, the amino acid substitution with the largest amount of associated bioactivity

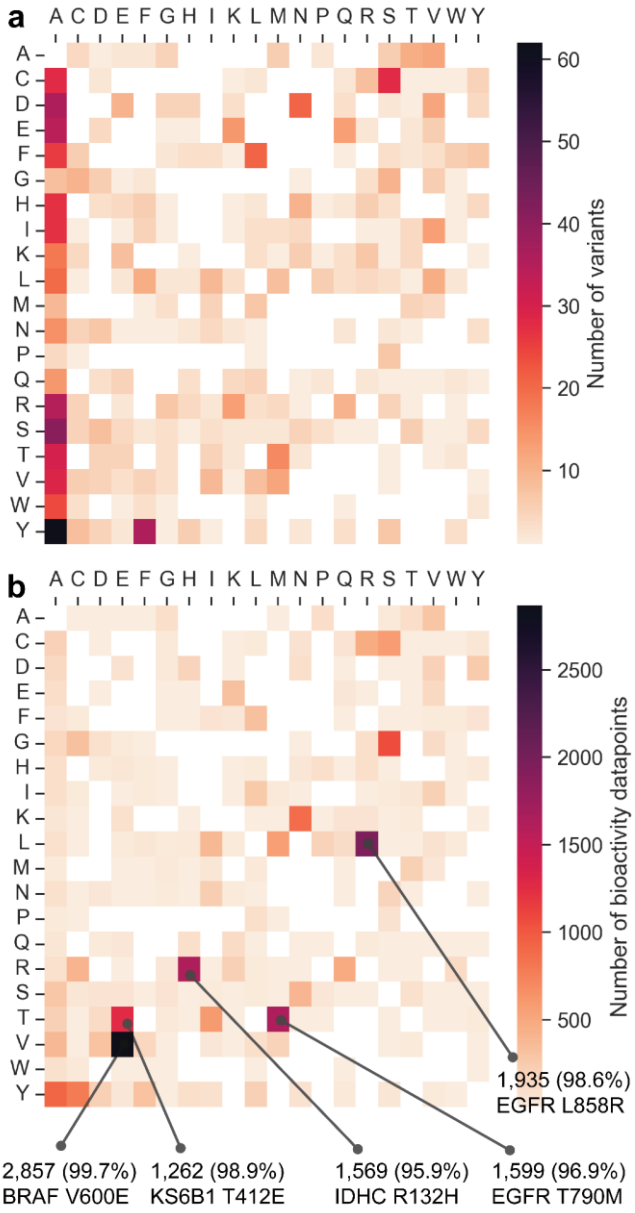

Figure 4. Amino acid substitutions reported in bioactivity databases. **a)** Unique variants reported per amino acid substitution. **b)** Number of bioactivity data points per amino acid substitution. Highlighted, is the substitution with the highest representation for the top five amino acid substitutions. In variants with multiple substitutions reported, each variant was accounted for individually.

data was valine to glutamic acid, with 2,864 bioactivity data points, out of which 99.7% corresponded to the BRAF V600E variant.

In line with the amount of data in ChEMBL per organism (**Supplementary Table 9**), the most frequently tested variants were in human proteins (BRAF, IDHC, RPKS6B1, EGFR). Indeed, out of the variant annotated bioactivity data, 90.56% corresponded to *Homo sapiens*. Viral and bacterial variants were also represented, however with only 4.82% and 0.70% of the bioactivity data. The remaining bioactivity data corresponded to 13 non-human Eukaryotic organisms of interest in preclinical studies, such as *Rattus norvegicus* or *Mus musculus*, among others.

The type of amino acid substitutions reported in bacterial variants were similar to human variants (**Figure 5a,b**). These featured many disruptive amino acid substitutions (91.53% in bacteria and 89.67% in human), either affecting the size or polarity of the original amino acid or, in most cases, both. To further characterise the disruptive potential of each amino acid substitution, we calculated the Epstein coefficient of difference[35], which is higher for more disruptive changes. In line with the previous observations, the Epstein coefficient of difference for most of the variants was higher than 0.4 (50.00% of the bacterial and 55.30% of the human variants), thus indicating changes in amino acid properties that would likely affect the protein's function. On the other hand, viral variants featured a larger proportion of conservative amino acid substitutions (17.81%, **Figure 5c**). This observation was also backed up by a lower proportion of amino acid substitutions with an Epstein coefficient of difference higher than 0.4 (41.21%), even when the size or polarity was affected. From a biological perspective, organisms with a higher mutation rate, such as viruses, are indeed prone to accumulate fewer damaging substitutions than organisms with a lower mutation rate subjected to more checkpoints, such as humans.

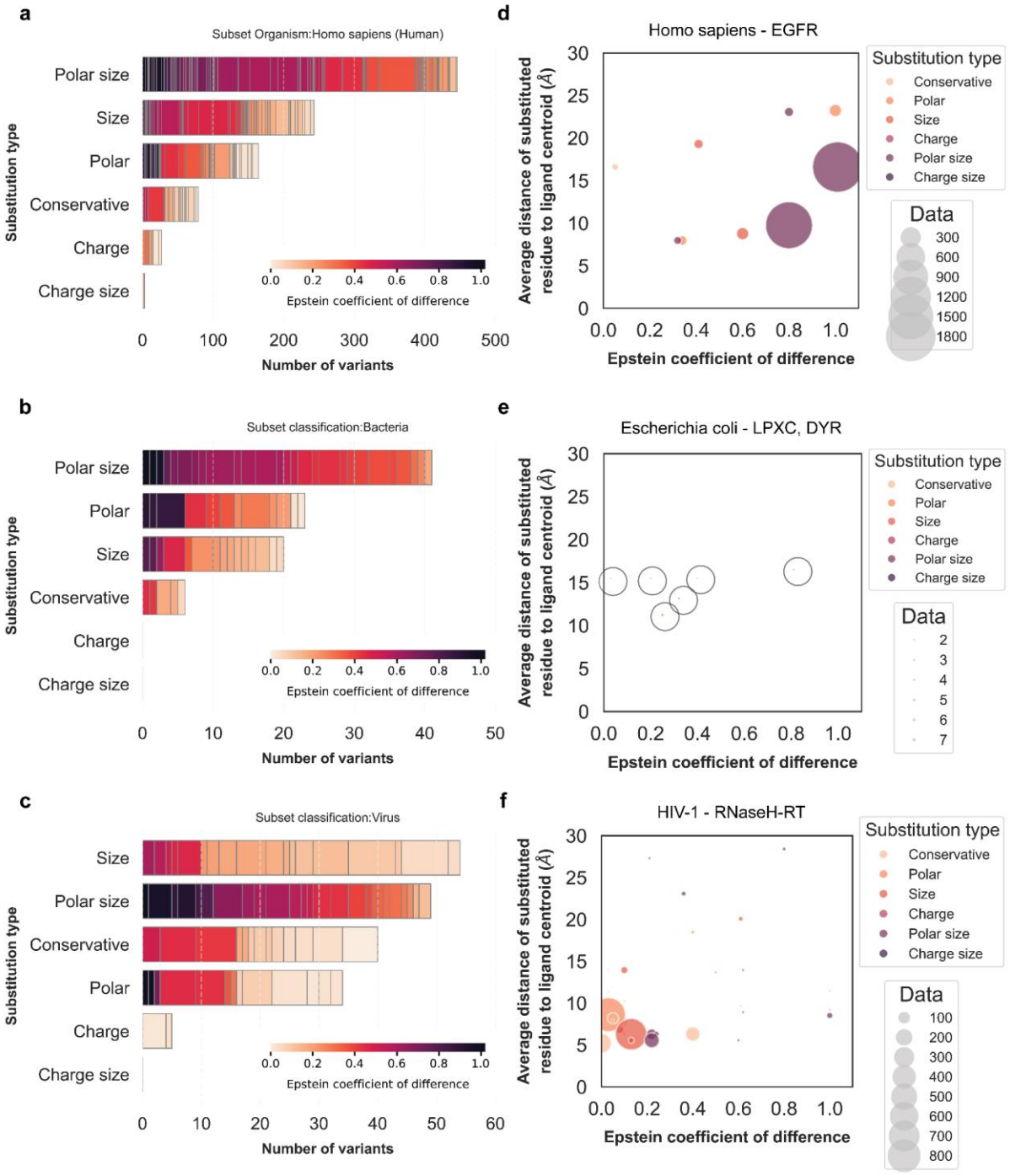

Figure 5. Types of amino acid substitutions in bioactivity databases across taxonomic categories: *Homo sapiens* (a,d), Bacteria (b,e), and Viruses (c,f). **a-c)** Number of variants according to their amino acid change, divided into six categories related to the effect in the amino acid polarity and size and coloured by the Epstein coefficient of difference of the corresponding amino acid substitution. **d-f)** Correlation between amino acid change relevance (Epstein coefficient of difference, x-axis), distance to ligand (average distance of substituted residue to ligand centre of geometry or centroid, y-axis), and sampling frequency (number of bioactivity data points, bubble size) in variants of **(d)** *Homo sapiens* EGFR (P00533), **(e)** *Escherichia coli* LPXC (P0A725) and DYR (P0ABQ4), and **(f)** *Human immunodeficiency virus* 1 (HIV-1) polyprotein RNase H - reverse transcriptase (RNaseH-RT, Q72547). Note that although Q72547 is the code for RNaseH-RT, the substitutions were concentrated in the RT domain, with only three substitutions in the RNaseH domain. In variants with multiple substitutions reported, each variant was accounted for individually.

Among the 14 viruses and 16 bacteria for which 217 and 115 variants were tested, respectively, two organisms concentrated the majority of the data available (**Supplementary Table 9**). HIV-1 accumulated 54.8% of the viral variants and 70.6% of the viral bioactivity data in just five proteins. Similarly, *Escherichia coli* concentrated 20.9% of the bacterial variants and 42.0% of the bacterial bioactivity data tested in eight proteins. A closer look into the nature of the substitutions reported in these organisms offered some interesting insights when compared to EGFR as a proxy for a human protein with disease-relevant variants. In line with the general observation across human proteins, the nine single substitutions reported for EGFR were few but of high relevance, with only one conservative substitution and Epstein coefficients of difference around (three) or higher than (five) 0.4 (**Figure 5d**). Based on the 77 crystal structures available, all reported EGFR substituted amino acids were located from 8Å to almost 25Å of the centre of geometry (centroid) of the protein ligands. Of note, the two most tested substitutions (resistance substitution T790M and activating substitution L858R) showed very high coefficients of difference but different locations with respect to the binding pocket (0.80 and 9.77Å, and 1.01 and 16.60Å, respectively). These two substituted residues are in the binding pocket of EGFR and correspond, respectively, to the gatekeeper residue and the back cleft. In contrast, HIV-1 RNaseH-RT harboured 31 single substitutions, of which 64.52% had an Epstein coefficient of difference lower than 0.4 (**Figure 5f**). Of note, these substitutions were concentrated around the non-nucleoside reverse transcriptase inhibitor (NNRTI) binding site, with distances to the ligand centroid mostly below 15Å. The only *E. coli* proteins with structural data, acetylglucosamine deacetylase LPXC (P0A725), and dihydrofolate reductase DYR (P0ABQ4), showed six substitutions affecting either size or polarity (**Figure 5e**), and were located around 15Å of the ligand centroid. The type of amino acid substitution, as well as the distance from the substituted residue to the ligand binding site, could affect the bioactivity of certain small molecules towards different variants. From a biological point of view, enriched human variants are likely to be disease-related whereas variants in pathogenic organisms are more likely linked to drug resistance. The extent of such an effect and its potential relevance in bioactivity modelling was analysed in the following sections.

Genetic variants affect bioactivity at different levels

Heterogeneity was found in annotated variants not only regarding the type and location of amino acid substitutions but also the number and structure of small molecules tested across them, as well as their relative bioactivity compared to WT. These observations reflected the interest in therapeutically targeting disease-relevant variants. In previous sections, it was shown that the majority of proteins have a small amount of variant bioactivity data compared to WT, in particular in proteins with sufficient data for modelling (**Figure 3**). Even in the proteins

with the highest variant bioactivity percentages (i.e. equal to or above 10%) that had sufficient data for bioactivity modelling (i.e. equal to or above 1,000 data), data density across variants was rather uneven. Out of the 13 data-rich proteins satisfying these conditions, WT was the most populated variant in all cases except for BRAF (P15056) V600E, IDHC (O75874) R132H, and RPKS6B1 (P23443) T412E (**Supplementary Table 10**), with the two first mutations corresponding to clinically relevant variants in cancer. BRAF and RPKS6B1 were also the only proteins, together with LRRK2 (Q5S007), where the most populated variant-annotated target had less than twice the amount of data of the second most populated variant, namely 1.52, 1.21, and 1.96 times. The rest of the proteins ranged from 4.73 (ALK, Q9UM73) to 104.64 (GNRHR, P30968) times more data in the most populated variant-annotated target – generally WT – compared to the second. The proteins with the largest relative data density differences between the first and second variants were those with the largest number of variants annotated (**Supplementary Figure 3a**). In these cases, the existence of many variants compensated for their data scarcity and still amounted to a relevant variant bioactivity percentage, above 10%. However, for all 13 data-rich proteins, only up to three variants – generally the most established clinically relevant – contained more than 500 data points, with some of the remaining variants dropping to as little data as one data point (**Supplementary Figure 3b**). These numbers corroborated the high data sparsity and hinted at the potential challenges to accurately reflect the differences in bioactivity caused by variants.

Two scenarios were contemplated to reduce the effect of chemical data sparsity across variants. The first one simulated an ideal scenario where all compounds would have been tested on all variants. For this purpose, fully dense common subsets were computed for targets with sufficient data, where only those compounds tested across all available variants were kept Given the number of variants with extremely low data density, this task was not trivial. In fact, approximately two-thirds of the 335 targets in the VEBD did not have a single compound that had been tested on all reported variants. For the other third consisting of 114 targets, the fully dense common subset represented a small portion of the target's set, with only 18 targets exceeding 10% and the maximum representation being 50%. Moreover, the size of their fully dense common subsets was very small, with only four targets surpassing 35 compounds tested across all their annotated variants (**Supplementary Figure 4**). However, the computation of fully dense common subsets proved to be relevant to achieve fair comparisons. In many cases, like for breakpoint cluster region protein BCR (P11274) and JAK2 kinase (O60674), the modelling protein set was highly biased towards WT bioactivity, making the fully dense common subset valuable for comparison (**Supplementary Figure 4b,c,f,g**). Given these results, a strategy was developed to compute non-fully dense common

subsets - referred to as common subsets - for the previously mentioned two-thirds of proteins for which a fully dense common subset was not available. Common subsets generated for compounds tested in at least two variants with a variant coverage of at least 20% identified 115 targets for which a fully dense common subset was not possible. Overall, using these parameters to compute the common subsets resulted in very diverse subsets covering 229 targets with an average common subset of 35 ± 121 unique compounds and 5 ± 6 variants. This was a clear improvement in terms of subset size from the original 114 fully dense common subsets, which had an average of 10 ± 33 unique compounds and 4 ± 7 variants. Additional measures were taken in very sparse targets by allowing the previous filters to be computed based on pairwise molecular similarity. This allowed us to include compounds only tested in one variant if a highly similar compound (e.g. Tanimoto similarity ≥ 0.80) had been tested in a different variant. The similarity option with the previously defined parameters allowed rescuing an additional four targets but did not improve the existing subset sizes, given the stringent 80% similarity threshold. The obtained similarity-expanded common subsets maintained the bioactivity distribution per variant of the VEBD, and all reached a higher balance and reduced sparsity as intended (**Supplementary Table 11**).

The generation of common subsets with varying parameters made it possible to analyse complete panels of compounds across variants. The versatility of such analysis on different protein families was exemplified for targets previously highlighted based on bioactivity data density and variant bioactivity percentage, namely EGFR (**Figure 6, Supplementary Figures 5,6**), HIV-1 RNaseH-RT, IDHC, and bromodomain-containing protein BRD4 - O60885 (**Supplementary Figures 7-9**, respectively). For EGFR, this analysis allows the user to follow some of the most biologically relevant activating – L858R, G719C/S, A750P, P753S, L861Q – and resistance – T790M – substitutions and the different generations of EGFR inhibitors (EGFRi) developed to achieve selective bioactivity profiles (as a reference commonly used in drug discovery we will consider a potency difference over 30-fold against specific variants of interest, which translates to a pChEMBL value difference over 1.5). The bioactivity analysis set for EGFR was generated from a common subset with compounds tested on at least three EGFR variants and variants covering at least 10% of the compounds. The analysis subset contained 22 compounds tested on nine out of the 14 annotated EGFR variants with clear differences in bioactivity (**Figure 6**, see **Supplementary Figure 5** for compound ID mapping). Out of these 22 compounds, 10 were approved drugs – EGFRi but also pan-kinase and other inhibitors – and the rest were either preclinical or clinical candidates (**Supplementary Figure 5,6**). The first two generations of EGFRi were represented in this analysis. First-generation EGFRi are reversible compounds developed to target activating mutations, in particular substitution L858R. Second-generation EGFRi are irreversible compounds aiming at a similar

selectivity profile. Three compounds (15-17), including second-generation EGFRi afatinib, showed consistently high pChEMBL values over 8.07**,** while seven (8-14) showed consistently low activity across variants with maximum pChEMBL value of 6.66.

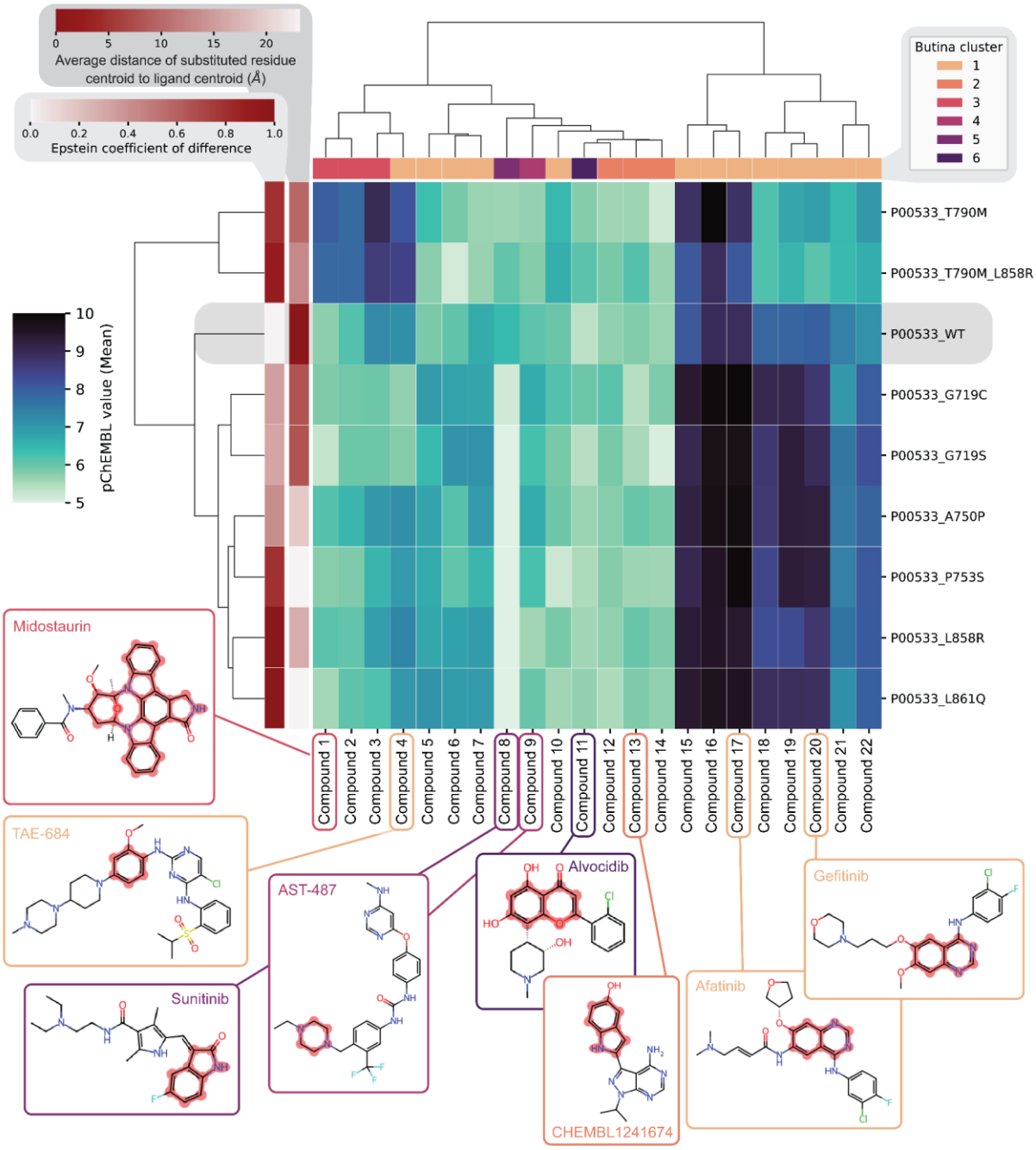

Figure 6. Full-panel bioactivity analysis of the effect of EGFR (P00533) variants. Bioactivity is represented in the heatmap as the pChEMBL value of different compounds, on the x-axis, tested on several variants, on the y-axis. See **Supplementary Figure 5** for the mapping of compound numbers to their connectivity ID, preferred name, and approval status. Compounds and variants were clustered by their overall bioactivity profile. Compounds are further represented by their corresponding Butina clusters upon clustering of the subset with a cutoff of 0.7. Compounds that are representatives of particular clusters or bioactivity profiles are highlighted and their 2D structures are displayed with the preferred molecule name (ChEMBL). The rest of the molecules can be found in **Supplementary Figure 6**. The biggest ring system in each molecule is highlighted in red for reference as a less stringent proxy

for the maximum common substructure to visually distinguish molecules with similar scaffolds. Variants are further represented by the distance from the substituted residue to the centroid of the ligand in the structure of the protein and by the Epstein coefficient of difference calculated for the amino acid substitution. In variants with multiple substitutions reported, the average distance and Epstein coefficient of difference are reported.

Moreover, four compounds (1-4) showed very high activity – pChEMBL value between 7.80 and 8.99 – against the two variants containing the resistance substitution T790M compared to the rest of the variants, including WT – where the maximum pChEMBL value was 7.34. These two variants, single substituted T790M and double substituted T790M/L858R, also exhibited the most different overall bioactivity patterns, as expected given their biological relevance. Indeed, five first-generation EGFRi (18-22) exhibited lower activity against the two T790M-containing variants (pChEMBL values between 6.09-7.00, compared to 7.01-9.33), as this resistance substitution is known to appear as a response to treatment with first- and second-generation EGFRi. Despite high activity overall, afatinib exclusively showed a decrease in bioactivity for the double mutant L858R/T790M. In terms of the location with respect to the ligand binding site, T790 is one of the closest substituted residues, below 10 Å from the ligand centroid and effectively in the binding site of EGFR. Additionally, the threonine to methionine amino acid change is highly disruptive with an Epstein coefficient of difference over 0.80. The rest of the variants behaved more similarly to the WT, with two major compound clusters with low (pChEMBL values between 5.00 and 7.34) and high activity (between 7.01 and 10.00), respectively. From these, WT was the odd one with the least marked differences between the two groups of compounds, as seen in the hierarchical clustering per variant (**Figure 6**). This was expected, as most EGFRi were developed to be variant-selective and reduce the side effects of anticancer therapies. The single substituted variant L858R behaved very differently from the double substituted T790M/L858R variant, in line with the different biological roles of these substitutions. Although the substitution to arginine is highly disruptive, L858 is further away from the ligand than T790. The Butina clustering performed on the 22 compounds showed that similar compounds exhibit similar effects across variants, as observed for clusters 2-6, and in line with the sequential development of EGFRi generations. Clusters 2-6 were populated by compounds with clear similarities, resulting in a diverse cluster 1 (**Supplementary Figure 6**) showing multiple patterns across variants but mostly containing first- and second-generation EGFRi. An interesting example was compound 4, which is structurally very different from the compounds in cluster 3 (compounds 1-3) yet exhibited the same bioactivity pattern. As such, this analysis can aid in the exploration of compounds with variant-selective profiles beyond the most well-known chemical groups. For other proteins, it can be a tool to rationalise the chemical modifications needed to develop drugs targeting specific resistance substitutions (**Supplementary Figure 7**); an instrument for extracting

starting scaffolds with specific selectivity profiles (**Supplementary Figure 8**); or to distinguish between compounds with different binding modes (**Supplementary Figure 9**).

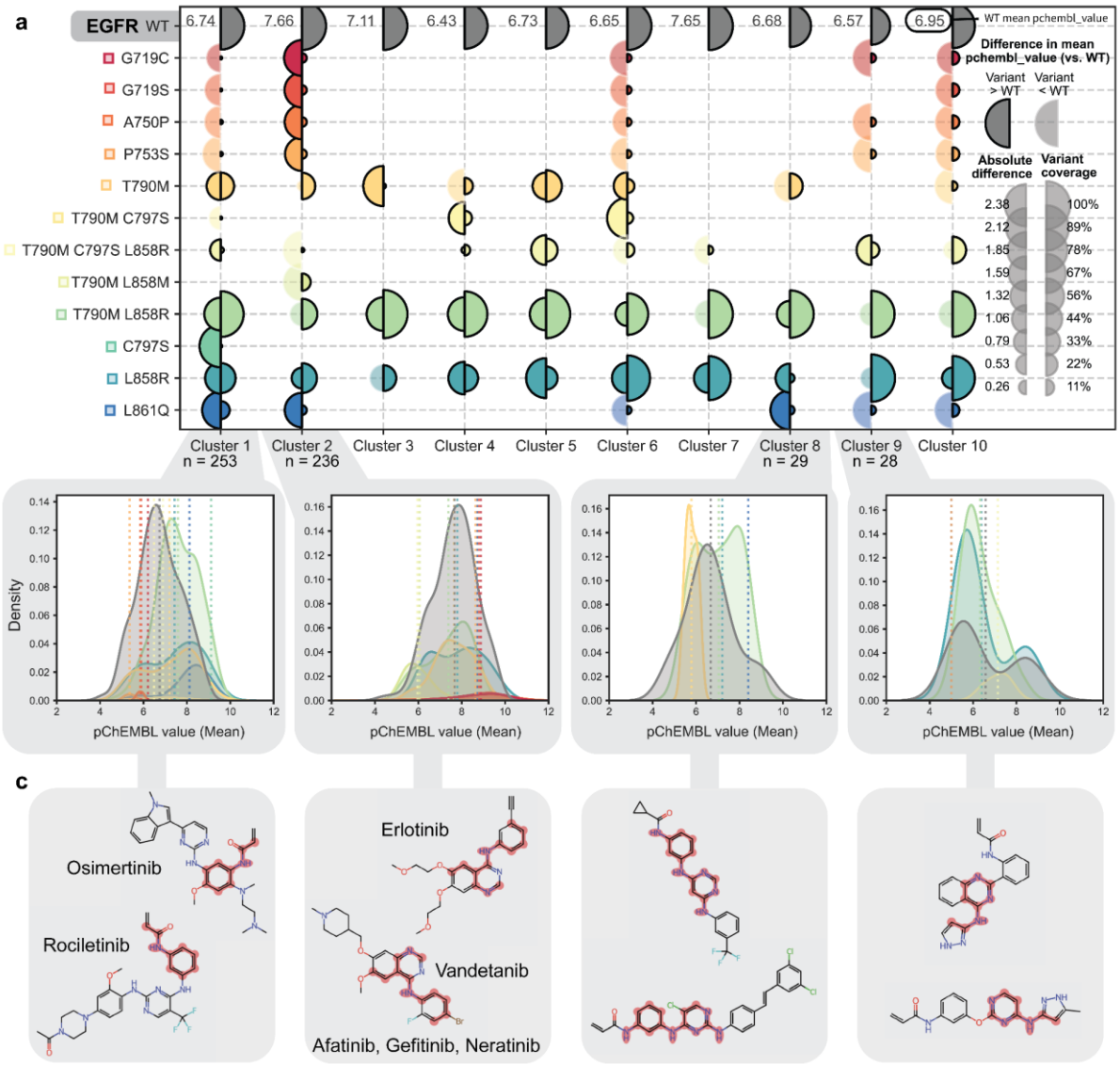

Figure 7. EGFR (P00533) bioactivity variability across variants compared to WT for compounds in the 10 most populated Butina Clusters upon clustering compounds tested on at least two variants with a clustering threshold of 0.5. **a)** Differences between mean *pchembl_value* in WT, displayed at the first row as calculated for the compounds in each cluster, and the mean *pchembl_value* in each of the variants for the compounds in the same clusters. The left bubbles represent the result of subtracting the variant mean from the WT mean. The bubble size represents the absolute value of this difference (error). Opaque left bubbles represent a positive error (i.e. the mean calculated for the variant is higher than for WT), and translucent left bubbles represent a negative error (i.e. the mean calculated for the variant is lower than for WT). Right bubble sizes represent the variant coverage, in other words, the percentage of compounds in each cluster that was tested on a specific variant. **b)** Distribution density of pChEMBL values across compounds in each cluster. Different colours represent the different variants where compounds of the cluster were tested, according to the colour code of panel a. Dashed lines represent the mean *pchembl_value*, which was used to calculate the differences in panel a. **c)** Two compound examples per cluster with the atoms corresponding to the maximum common substructure of all the compounds in the cluster highlighted in red. When available, approved compounds or preclinical candidates are displayed.

The different effects observed for different chemical clusters in common subsets could also be expanded to bigger yet sparser subsets. This allowed us to analyse the overall effect of variants on different subsets of the chemical space tested for one protein. While this analysis is possible for the whole protein subset, in targets with a clear bias towards WT testing, selecting subsets of compounds tested on at least two variants was still preferred to increase the significance of comparisons across variants. Particularly for EGFR, the set of 1,219 compounds tested on at least two variants was clustered using the Butina algorithm[36] with a threshold of 0.5 resulting in 118 clusters. Clear differences in bioactivity across variants were observed among the top 10 biggest clusters (**Figure 7**). Chemistry-related changes in bioactivity distribution were already somewhat apparent on the WT level (**Figure 7a,b**), with mean pChEMBL values between 6.43 and 7.66 from slightly divergent distributions. The compounds in the two most populated clusters (n=253 and n=236, respectively) were tested across 11 and 10 out of the 12 variants, respectively, with various rates of variant coverage (**Figure 7a**). These two clusters included approved first (cluster 2), second (cluster 2), and third generation (cluster 1) EGFRis, as well as pan-kinase inhibitors (cluster 2). Third-generation EGFRis were not present in **Figure 6** and were developed to selectively target the L858R/T790M double substitution. Furthermore, the average differences in bioactivity compared to WT across variants were virtually the opposite between the two clusters, in line with the known selectivity profiles of different generations of EGFRi. For example, compounds tested on rare variants G719C, G719S, A750P, and P753S all showed lower activity than compounds tested on the WT in cluster 1 (0.54, 0.85, and 0.88 points below WT – 6.74) but higher in cluster 2 (1.21, 1.11, and 1.06 points above WT – 7.66). The opposite effect was observed for compounds tested on the double substituted T790M/L858R variant, which had a mean pChEMBL value 0.85 points higher than compounds tested on the WT in cluster 1 (7.59 vs. 6.74) and 0.27 points lower than compounds tested on the WT in cluster 2 (7.39 vs. 7.66). Of note, the bioactivity distributions across compounds tested in each variant were highly diverse (**Figure 7b**), thus relevant in addition to the point mean differences. Together, this type of analysis pinpoints chemical patterns (as highlighted in **Figure 7c** for the maximum common substructures of compounds in each cluster) driving differences in bioactivity across variants. Similarly to EGFR, this analysis can help expand the results observed in the full-panel bioactivity analysis for other proteins as exemplified for HIV-1 RNaseH-RT, IDHC, and BRD4 (**Supplementary Figures 10-12**, respectively). In an explorative fashion, results derived from this analysis can be the starting point of drug design campaigns satisfying certain activity characteristics. Alternatively, in virtual screening campaigns, they can be relevant for decision-making to reduce noise in models or increase the modelling performance by constructing variant-aware models, as explored in the following section.

Variant awareness improves modelling performance

The effects of variant bioactivity data on the performance of machine learning modelling were investigated by comparing results obtained from three scenarios. The first scenario corresponds to modelling in a variant-agnostic situation, wherein all bioactivity data measurements are (mistakenly) assumed to derive from assays carried out on WT proteins only (*QSAR-All*). The two other scenarios correspond to modelling in a variant-aware situation, wherein data points assayed on variant targets are either kept in (*PCM-All*) or filtered out of the training set (*QSAR-WT*).

First, modelling performance was evaluated based on random split cross-validation on the VEBD in its entirety, splitting out each protein in turn, to assess the overall effect of introducing variant-aware strategies. As expected, on average the performance of models decreased with a scarcer number of bioactivity data points (**Table 1, Supplementary Figure 8a and 8c, and Supplementary Table 12**), characterised by the average cross-validated Pearson correlation coefficient (Pearson's r) below 0.40 when modelling proteins with 5 to 100 data points, around 0.70 with 100 to 500 data points, around 0.75 with 500 to 200 data points, and above 0.76 with more than 2000 data points, respectively. In any case, variant-aware models showed increased performance, with all '*QSAR-WT'* models showing an increased correlation with experimental values compared to '*QSAR-All*' models. Data balance between the data points obtained on WT and the ones on variant targets had an impact on the significance of the differences in performance observed (**Table 1 and Supplementary Figure 13b and 13d**). This was demonstrated in protein families with substantial experimental data by the significantly increased performance of the `*PCM-All*` model (0.716) for protein kinases (p-value=4.1 x10^{-5}), with 9.1% of variant bioactivity percentage, compared to that of `*QSAR-All*` and `*QSAR-WT*` models (0.700 and 0.701, respectively). In contrast, no significant difference was observed for family A GPCRs, ion channels, and nuclear receptors, which all had a lower data balance (between 1.5 and 2.5% variant bioactivity percentage), and for which PCM was not the best strategy. Indeed, all points relating to protein kinases in **Figure 8a zoom-in** were very close to or below the identity line, and most data-rich kinases showed a significant performance increase when using PCM. These included EGFR (P00533), ABL1 (P00519), LRKK2 (Q5S007), ALK (Q9UM73), RPKS6B1 (P23443) and proto-oncogene RET (P07949) kinases, all having more than 2,000 associated data points with at least 10% variant bioactivity percentage, with correlation coefficients between 0.75 and 0.85 (**Table 1, Figure 8a**). Interestingly, of data-rich proteins, only BRAF (P15056) showed a decreased performance when including data points of variants, with a Pearson's r of 0.847 for PCM-All and 0.858 for QSAR-WT. This could be the result of the very large amount of data points associated with

variants (60.3%) and due to the distinctively divergent but overlapping trends in the distributions of bioactivities between WT and variants (**Supplementary Figure 4**). These results highlight the importance of variant awareness in bioactivity modelling but do not provide a solid basis for general recommendations on the variant-aware strategy that should be used.

Table 1. Modelling performance of variant-annotated proteins following three modelling strategies: PCM explicitly modelling variants (PCM-All), QSAR with all protein data without considering variants (QSAR-All), and QSAR removing variant data (QSAR-WT). The performance of PCM and QSAR models depends on the number of data points and the variant bioactivity percentage. Performance is reported for the entire training set, focused protein families, and data-rich proteins (more than 1,000 data points with at least 10% variant bioactivity percentage) for a random split 5-fold cross-validation strategy as the average Pearson correlation coefficient for each group or protein and, between brackets, as the average per group or protein of the standard deviation of Pearson r between cross-validation folds for each protein. The best average Pearson r is reported in bold for each row. Pearson r of PCM and/or QSAR-WT models significantly differing from QSAR-All models are starred. Pearson r of PCM or QSAR-WT models significantly differing from all other models (i.e. QSAR-WT and QSAR-All, and QSAR-All and PCM-All respectively) are underlined. Statistical results are detailed in **Supplementary Table 17**.

	Average Pearson correlation coefficient (average standard deviations)			# data points	variant bioactivity (%)
	PCM-All	QSAR-All	QSAR-WT		
All	0.653 (0.117)*	0.634 (0.116)	**0.654 (0.121)***	453,660	5.5
5 to 100 data points	**0.396 (0.322)**	0.352 (0.323)	0.363 (0.378)	3,257	29.1
100 to 500 data points	**0.704 (0.085)**	0.690 (0.083)	0.691 (0.094)	19,694	10.0
500 to 2,000 data points	0.746 (0.038)	0.737 (0.039)	**0.747 (0.041)***	84,426	4.5
2,000 to 20,000 data points	**0.769 (0.018)***	0.763 (0.017)	0.764 (0.017)	346,283	5.2
Family A GPCRs	0.731 (0.046)	0.735 (0.035)	**0.752 (0.037)**	93,454	1.8
Ion Channels	0.620 (0.142)	0.613 (0.134)	**0.646 (0.168)**	16,635	1.5
Nuclear Receptors	0.704 (0.047)	0.690 (0.036)	**0.714 (0.034)**	14,344	2.5
Protein Kinases	**0.716 (0.068)***	0.701 (0.068)	0.700 (0.080)*	133,396	9.1
P00533 (EGFR)	**0.822 (0.009)***	0.802 (0.008)	0.809 (0.004)	13,601	18.4
Q72547 (HIV-1 RNaseH- RT)	**0.809 (0.013)***	0.764 (0.005)	0.776 (0.012)	6,953	34.0
P00519 (ABL1)	**0.867 (0.008)**	0.850 (0.019)	0.857 (0.012)	4,985	22.3
P15056 (BRAF)	0.847 (0.012)	0.834 (0.013)	**0.858 (0.014)**	4,740	60.3
P36888 (FLT3)	**0.813 (0.022)**	0.812 (0.016)	0.798 (0.018)	4,390	11.8
O60885 (BRD4)	0.856 (0.007)*	0.714 (0.038)	**0.858 (0.013)**	4,106	17.1
P10721 (KIT)	**0.748 (0.028)***	0.708 (0.010)	0.716 (0.015)	2,897	19.4
Q5S007 (LRRK2)	**0.853 (0.017)**	0.851 (0.013)	0.827 (0.009)	2,760	34.0
Q9UM73 (ALK)	**0.854 (0.017)**	0.829 (0.011)	0.837 (0.021)	2,598	24.9
P23443 (RPKS6B1)	**0.854 (0.005)**	0.853 (0.012)	0.682 (0.042)*	2,286	55.2
O75874 (IDHC)	**0.804 (0.014)**	0.759 (0.031)	0.775 (0.045)	2,203	86.3
P07949 (RET)	**0.778 (0.027)***	0.752 (0.033)	0.718 (0.020)	2,123	13.2
P30968 (GNRHR)	**0.758 (0.047)**	0.724 (0.030)	0.720 (0.045)	1,921	23.7

Next, the capacity of models to predict the bioactivity of compounds on unseen variants was investigated. To this end, Leave-One-Variant-Out (LOVO) cross-validation was carried out. This confirmed the trend previously observed of the ability of PCM models to interpolate in the protein feature space, especially for richer sets of proteins (more than 2000 data points) with an average Pearson's r of 0.325 compared to 0.311 for other proteins (**Supplementary Table 13**). To decrease the sparsity of datasets, similarity-expanded common subsets were derived to focus on a subset of molecules and their analogues tested across a subset of variants. The latter drastically decreased the applicability domains of models (**Supplementary Table 14**) and affected the performance of most models (**Figure 8b and Supplementary Table 15**) but improved the performance of models when used in combination with LOVO cross-validation (**Figure 8c and Supplementary Table 16**) for most proteins. Nonetheless, the general trend showed no clear difference between '*QSAR-All*' and '*PCM-All*' models derived from LOVO cross-validation from the common subsets (**Figure 8d**), suggesting that the extrapolation to new variants using PCM is similar to random prediction. These results show the complexity of accurately predicting bioactivity for individual variants. Moreover, they highlight the impact of data sparsity on model performance and how the limited size of current datasets restricts extrapolation in the protein feature space when focusing on analogue molecules.

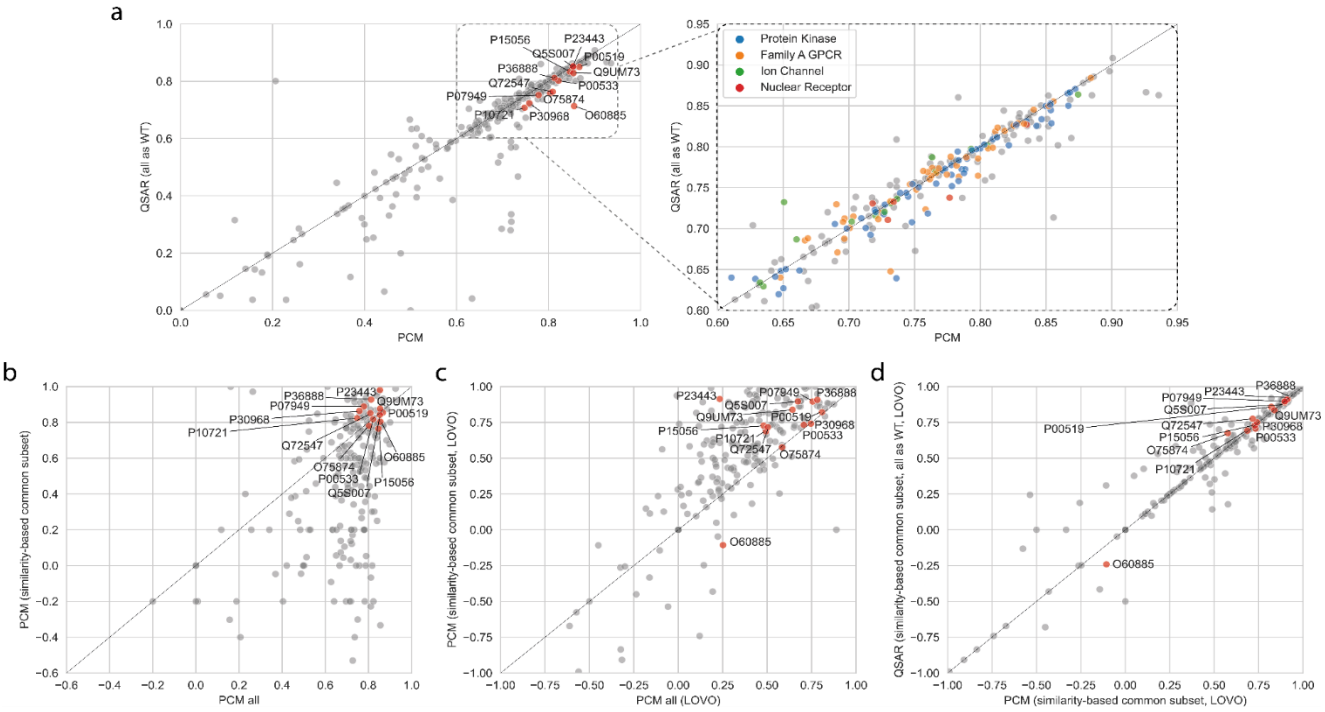

Figure 8. Comparison of the performance (cross-validated Pearson correlation coefficient average) of models in a variant-aware (PCM-All) and variant-agnostic (QSAR-All) setting considering the newly annotated set and zoom-in coloured by of protein families (a). Comparison of performance between the 'PCM-All' and 'QSAR-All' models on the similarity-based common subset (b) and the newly annotated set using Leave-One-Variant-Out (LOVO) cross-validation (c). Labelled points correspond to data-rich proteins (see **Table 1**).

The general trends highlighted above were consistent across the data-rich proteins, although few of them had a significant performance improvement when using variant-aware models (**Table 1**, **Figure 8**). On a protein-specific level, this effect can be traced back to data sparsity and imbalance across variants and subsets of the chemical space (**Figure 7** and **Supplementary Figure 10** for EGFR and HIV-1 RNaseH-RT, respectively). In fact, tackling these issues by reducing the applicability domain with a similarity-expanded common subset resulted in equivalent or improved PCM performance in random split cross-validation compared to complete sets for these proteins, with a clear advantage over the variant-agnostic model (**Supplementary Table 13, Figure 8b**). Moreover, the analysis of the bioactivity patterns can help explain discrepancies from the general modelling trend. For example, among data-rich proteins, BRD4 (O60885) displayed the biggest increase in performance when using variant-aware models in random split cross-validation (**Figure 8a**). Following the general trend, we expected a good extrapolation to novel variants for this protein, which was not the case (**Figure 8c**). The examination of the substituted residue distance to the ligand's centroid on the bioactivity cluster map for BRD4 (**Supplementary Figure 9**) highlighted that the two most represented variants, Y97A and Y390A respectively, are each part of different protein domains, bromodomains 1 and 2 respectively, corresponding to different binding sites, and had therefore opposite effects on bioactivity for the subset of compounds examined. This was confirmed in the protein's structure and explained the lack of generalization power of the model, which might be improved by splitting the chemical space into domain-specific binders. Still looking at the data-rich proteins, IDHC (O75874) showed poor extrapolation, which could be traced back to the very similar bioactivity profiles across the tested variants, all of them occurring in the clinically-relevant R132 residue (**Supplementary Figures 8,11**). Based on this information, model performance could be improved by pooling all variant data or designing protein descriptors able to capture the subtle differences in one residue. These results stress the importance of informed decision-making via the analysis of bioactivity trends to design relevant training sets and strategies for variant-aware modelling.

Discussion

Bioactivity modelling is one of the cornerstones of computational drug discovery. Despite the most recent advances in modelling techniques and capacities, data quality and quantity remain a major bottleneck, particularly for those working in the public sector without access to large proprietary or commercial datasets. As a consequence, large, curated, and open bioactivity databases such as the ChEMBL database or the Papyrus dataset constitute key resources for the community. Despite the many benefits that the expert extraction and curation processes for these databases provide, the user still needs to navigate the often-complex

database structures and make informed decisions to select and curate data for the modelling task at hand. This does of course also reflect the fact that developing, running and processing the data from bioactivity assays is a complex scientific endeavour. Careful selection of several fields in these databases, such as activity comments and assay types can have a big impact on the quality of the modelling data. Here, the effect of a commonly overlooked field in bioactivity databases, amino acid substitutions constituting protein variants, was extensively analysed. The genetic variability landscape in the ChEMBL database has been explored in detail here for the first time, including the annotation strategy, the extent of variant data at different levels, the effect on bioactivity distributions, and finally the effect on bioactivity modelling. The dataset and results from this are made available to facilitate modelling with consideration of genetic variants. Moreover, a full analysis Python package is made available at https://github.com/CDDLeiden/chembl_variants to promote variant analysis in proteins of interest to the user and thus help make informed decisions about data selection and curation for modelling.

A variant annotation strategy parallel to that of ChEMBL was developed that extracted 82.65% of the original variant annotations from the assay descriptions, which reinforced the confidence in the original ChEMBL variant annotation pipeline (which delivers these annotations by manual extraction of protein variant information from original papers). A clear advantage in the ChEMBL pipeline was the access to expert knowledge to rescue variants otherwise missed by a regular expression match. For example, sequence number shifts and non-canonical amino acid substitution definitions were identified among these expert rescues. However, misannotations reported by ChEMBL were also identified, for example, derived from mistakenly linking assays to protein families rather than single proteins. The current annotation strategy also retrieved several substitutions that had not been previously reported in ChEMBL 31. Nevertheless, these results need to be considered cautiously since they are based on fields previously extracted by ChEMBL rather than the original source in the literature and might miss important aspects of the experimental set-up. Importantly, this approach also relies on accurate reporting of tested variants in the scientific literature in order for their subsequent capture in bioactivity databases. Collaborative work such as reported here is key to improving the ChEMBL database[37,38] for the wider community; for future releases of ChEMBL we will aim to improve and enhance our reporting of variant data based upon the findings in this paper . Although several drug and protein databases contain variant data, the effect of drugs on specific variants is very sparse and conflicting[39,40]. An expert-curated dataset derived from our analysis could therefore serve as a user-friendly central repository for variant bioactivity data regularly retrieved from ChEMBL and additional sources. As a result of this collaboration, a revised version of this work will be released, integrating the alterations recommended through

the feedback loop (see ChEMBL comments on **Supplementary Table 1** and **2**, revision ongoing).

The variant landscape in ChEMBL 31 and additional Papyrus sources is, as expected, a reflection of the clinical relevance and interest of the community in particular organisms, protein families, targets, and individual variants. Unsurprisingly, human proteins concentrated the bulk of the variant data, but several mammalian orthologs and human pathogens were also identified. Of note, curated drug resistance databases for significant pathogens such as HIV[41], tuberculosis[42], and other antibiotic-resistant bacteria[43] are available independently of bioactivity databases and should be queried separately. Apart from being more complete, these databases have a more domain-focused curation process e.g. strain annotation in microorganisms. Although different organisms show significant differences in the amounts of data available, the amino acid substitution trends align with nature-observed patterns. Indeed, organisms with smaller genome sizes and higher mutation rates, such as viruses and to a lesser extent bacteria, accumulated larger amounts of non-disrupting substitutions compared to human proteins[44,45].

Among human protein families, enzymes, in particular kinases, amassed the most variant data, though not always proportionate to the overall data volume. While these numbers do not correspond to evolutionary mutation rates[46], they are certainly correlated to the high interest in protein kinase variants in cancer research[47]. Indeed, the targets that simultaneously displayed high variant bioactivity percentages and large amounts of data overall were predominantly cancer-related kinases with clinically relevant somatic substitutions such as EGFR[48], ABL1[49], BRAF[50], and ALK[51]. Nonetheless, in this category were also cancer-related kinases with no reported disease-related somatic substitutions like RPKS6B1[52], where experimental mutations are common, or kinases responsible for other pathologies, such as LRRK2 in Parkinson's[53]. Of note, the individual variants reported for specific targets also reflect the interest within the scientific community and do not necessarily include all reported and clinically relevant variants[54]. Other than clinically relevant variants, experimentally important variants were found, such as activating substitutions in downstream cascades[55], or alanine scanning panels for functional[56], or thermostabilizing assessment[57] in GPCRs. Far from negligible, such panels can be repurposed for model training, consequently reducing the need for experimental assays[58].

The Python package and notebooks that accompany this work have been carefully designed to allow complete reproducibility of the annotation and variant landscape analysis. However, their primary purpose is to empower readers to self-assess variant effects on protein bioactivity. As shown here for the clinically relevant kinase EGFR, among other data-rich

targets, these analyses can identify clusters of chemical space with varying effects on bioactivity, specific protein structural traits causing differing bioactivity patterns, and compounds with desirable selectivity profiles. These results not only are in line with the literature and enabled the analysis of activating and resistance-inducing substitutions, but also extended beyond the most widely-recognized variants and chemical classes[59]. In turn, they can be used as hypothesis generators in drug design[60] as well as recommendation systems to include or remove certain chemical clusters[61] or variants from a prospective modelling or virtual screening task[62]. Indeed, for a target like EGFR with a high variant bioactivity percentage and differential bioactivity profiles across variants and chemical groups, our bioactivity modelling results indicated a decrease in predictive performance when variants were not accounted for, generalizing the effects previously observed when modelling cyclooxygenases 1 and 2[63]. Both removing variant data from the QSAR model and explicitly modelling each variant in a PCM model increased performance in random split cross-validation, likely by reducing the negative effect of noise[64,65]. Similar results were observed for other proteins with a high variant bioactivity percentage despite large inter-target variability. Nevertheless, non-optimized protein sequence descriptors were used in this work. Furthermore, the average length of protein sequences varies greatly - for instance considering the 566 amino acids of HIV-1 RNaseH-RT and the 2549 amino acids of the human mammalian target of rapamycin (MTOR) - and could influence the sensitivity significantly and hence the ability of PCM to detect signal from the averaged representation used herein. To remedy these challenges, the use of alignment-dependant or autocorrelation descriptors could be explored[8,66]. Moreover, as previously mentioned, some mutants are disease-causing and are often the drug target. For these cases, in which molecules are optimised away from the WT, the baseline for the QSAR-WT could be substituted with the disease-causing mutant.

The modelling results presented here for all proteins containing variant data can be used for decision-making regarding additional data curation or the selection of modelling tasks for individual proteins. As a rule of thumb, targets with small datasets and/or high variant bioactivity percentages are the most susceptible to the presence of variants. These should be thoroughly examined before modelling and, if needed, additional measures should be implemented to tackle the drawbacks in the dataset[67].

Beyond bioactivity modelling with a focus on the WT protein, the dataset and results presented here can be exploited in variant bioactivity prediction with some precautions. First, variant data is still too sparse for large-scale modelling of new variants, as represented by the low performance of PCM models with LOVO validation. However, small-scale campaigns following data balancing strategies showed promising results and should be considered in light of each

particular project's scope[68]. Second, in this work only amino acid substitutions were considered, however, other aberrations such as deletions, insertions, amplifications, or copy number variations are known to be clinically relevant and affect both protein function and pharmacology[48,69]. A protocol should therefore be devised to also map these variations in bioactivity databases accurately. Third, the biological context of the variants studied – activating vs. resistance substitutions, as an example – is correlated with the effect in bioactivity, and should be considered in database annotation and extrapolated to modelling. Fourth, new clinical variants are constantly identified and have limited data in bioactivity databases compared to established variants[70]. This does not mean that these variants are less important, and thus more appropriate channels for variant tracking should be consulted simultaneously to assess clinical relevance. Finally, the data and results presented here should not be restricted to bioactivity modelling for virtual screening, and thus the exploration of other modelling tasks considering protein variants is highly encouraged including (and not restricted to) selectivity modelling[71], drug design by fragment merging[72], or pharmacophore modelling[62].

Conclusion

The genetic variability landscape of ChEMBL, the most widely used public bioactivity database in computational drug discovery, was comprehensively analysed for the first time. Key advantages resulting from years of expert knowledge gathering in ChEMBL's variant annotation pipeline were identified through parallel annotation. Additionally, misannotations requiring future correction were found. Recommendations for pipeline enhancement were provided, alongside a proposal for simplified annotation of target variants for bioactivity modelling, which are made available in a modelling dataset. The amount and distribution of variant data across protein organisms, families, individual proteins, and variants were extensively described. Furthermore, a Python package and notebooks were developed to assess variant effects on bioactivity distributions and modelling performance. The potential of these analysis tools to extract variants and promising chemical candidates was demonstrated, particularly for data-rich proteins. Particularly, informed decisions for noise reduction in bioactivity models and modelling variant bioactivity can be facilitated using our approach.

Materials and methods

Bioactivity data sources

Bioactivity data was collected from ChEMBL (version 31) and the Papyrus dataset (version 5.5). The Papyrus dataset contains highly curated data from ChEMBL version 31, ExCAPE-DB, and other individual datasets. Protein targets in the Papyrus dataset are identified either by *accession* (i.e. UniProt accession code) or *target_id*. The latter is constructed from the accession and the amino acid substitutions present in the variant analysed, with *accession_WT* for wild-type (WT) proteins. In its current version, the Papyrus dataset does not reflect variants described in ChEMBL.

ChEMBL data was collected using the ChEMBL Python client (**Supplementary Figure 14a,** full query available on the associated GitHub repository). The data queried included activities (i.e. *pchembl_value* and *activity_comment*), assay descriptions, molecular structures (i.e. SMILES – *canonical_smiles*), protein identifiers and sequences, and ChEMBL-annotated variants (i.e. *mutation* in the *variant_sequences* table).

After assay-based amino acid substitution annotation (see "Amino acid substitution annotation" section and **Figure 1**), ChEMBL assay-target pairs were given Papyrus-like identifiers based on the validated substitutions. Target variants were henceforward identified by *target_id*. Subsequently, individual ChEMBL activity points were mapped to annotated variant targets (*target_id*) based on their *assay_id* and *accession*. Duplicated activity data (*target_id-compound chembl_id* pairs) from several assays were joined into one single point by dropping low-quality data and calculating the mean pChEMBL value or most common activity label (**Supplementary Figure 14b**). The data_validity_*comment* field was used to drop low-quality data (author confirmed error), as done in the Papyrus dataset.[34] The activity_comment field was also used to define active and inactive binary labels when *pchembl_value* was not available.

Before variant bioactivity analysis, the Papyrus and ChEMBL datasets were integrated. Firstly, only the Papyrus entries originating from the Christmann subset were considered, filtering out de facto any Papyrus data point with ChEMBL as a source, avoiding duplicates. ChEMBL compounds were given Papyrus-like identifiers (*connectivity*). Then, the average *pchembl_value* was calculated for unique *target_id-connectivity* pairs. For data points with no *pchembl_value*, the most common activity label was kept. Finally, the VEBD for analysis was constrained to only targets with at least one variant annotated other than the WT.

Amino acid substitution annotation

ChEMBL amino acid substitutions were extracted from assay *descriptions* for unique assay-target (i.e. *assay_id-accession)* pairs following a three-step approach (**Figure 1**).

i) First, regular expressions were used to extract from the assay description amino acid substitution patterns. This is, either a one-letter amino acid code followed by an unlimited number of digits and another one-letter code, or a three-letter amino acid code followed by digits and another three-letter code. Subsequently, three-letter codes were transformed into one-letter codes.

ii) Second, exceptions were defined from assay-associated metadata and filtered out. These exceptions included assay cell types, target names, and target gene names and synonyms. At this level, an option was included to manually define exceptions from a JSON file for specific assays. Here, most "M1" and "D2" instances were filtered out as they could easily get a false positive validation status in step iii. The complete JSON file used for manual exception definition is included in the associated GitHub repository.

iii) Third, the remaining substitutions were validated by mapping the first amino acid of the substitution pattern to the WT sequence. If the mapping was successful, the substitutions were included for further analysis.

The resulting annotated assay-target pairs from the first round of annotation were introduced in an annotation feedback loop where they were compared to the original ChEMBL-annotated variants (**Supplementary Figure 1**). Annotations missed by ChEMBL were manually checked to assess their validity and classified accordingly into different categories of true and false positives. True positives included likely correct new annotations and likely correct rescue instances of “UNDEFINED MUTATION” labels in ChEMBL. New annotations and rescues with deletions were also categorized as true positives given the scope of this work. ChEMBL-only annotations were parsed and categorized into different categories of true and false negatives. True negatives included misclassified annotations due to the mislinking of single protein assays to protein families. Missed deletions were also categorized as true negatives in light of this work’s scope. False negatives included instances where expert knowledge was required. These were, for example, variants for which the amino acid substitution extracted matched but the sequence position was different due to sequence number shifts. Another example was constituted by completely missed substitutions because they did not correspond to the canonical regular expression. On the verge between true and false negatives were other ambiguous sequence number and amino acid substitution mismatches that did not correspond to the categories defined before. Without further manual curation, these could correspond

either to potential ChEMBL miss-annotations or missed correct annotations requiring expert knowledge. In a second round of annotation following the annotation feedback loop, the defined false positives were excluded from the annotated variants and reverted to WT. Similarly, false negatives were rescued by using the ChEMBL-annotated variants. The ambiguous cases were annotated as undefined variants given the lower confidence. The assay-target annotations from the second round were further linked to ChEMBL activity data to annotate variant targets (see section “Bioactivity data sources”).

Family and taxonomic distribution analysis

Protein family annotations were retrieved from ChEMBL version 31 by querying levels L1-L5 from the SQL table protein_family_classification for all unique UniProt accession codes. Proteins in the VEBD were mapped to their corresponding family levels based on their accession code. Non-defined levels were labelled as ‘Other’. On levels L1 and L2, small-sized families were grouped into larger families as follows. L1 tags ‘Auxiliary’, ‘Unclassified’, ‘Structural’, and ‘Surface’ were grouped into ‘Other’. L2 tags ‘Primary active’, ‘Ligase’, ‘Isomerase’, and ‘Writer’ were grouped into ‘Other’. Additionally, all G protein-coupled receptor L2 tags were grouped into a single L2 family, ‘GPCR’.

Subsequently, the total number of bioactivity data points as well as the number of variant bioactivity data points in the VEBD were calculated across families for each level. From these, the variant bioactivity percentage per family was calculated by dividing the amount of variant data by the amount of total data and multiplying the result by 100. Similarly, the novel variant bioactivity annotation percentage was calculated exclusively in ChEMBL data by dividing the number of bioactivity data points in potentially novel annotated variants (i.e. not previously defined in the ChEMBL ‘mutatation’ variable) by the total number of variant bioactivity data and multiplying the result by 100.

Organism names and HGNC gene symbols were mapped on accession codes from the Papyrus version 05.5 protein table. Moreover, the proteins’ taxonomy was retrieved and mapped for all unique UniProt accession codes using the UniProt API via the UniProtMapper package. The two *Escherichia coli* strains present in the dataset were aggregated under one single *Escherichia coli* organism. The number of variants and bioactivity data points were subsequently calculated at different taxonomy levels.

Statistical analysis per protein and variant

The amount and distribution of variant bioactivity data across individual proteins and variants were analysed in detail. For each protein, the number of variants and bioactivity data points

were calculated, as well as the variant bioactivity percentage compared to the totality of the protein's data. Within proteins, variants were ordered from most to least populated in terms of bioactivity data. The relative amount of data in the most populated compared to each of the following variants was calculated by dividing the amount of data in the first variant by the amount of data in the variant of interest.

Amino acid substitution type analysis

Amino acid substitution types were extracted from the variants. For variants with multiple substitutions, all the substitutions were considered individually. Three substitution-type definitions were implemented:

i) Categorical: Six substitution-type categories were defined based on the type of amino acid substitution regarding side chain size and polarity. 'Conservative' for amino acid substitutions where the size and polarity remained similar. 'Size' when size changed but polarity remained the same. 'Polar' and 'Charge' when the size remained similar but either the polarity or the actual charge, respectively, changed. And 'Polar size' and 'Charge size' as a combination of the aforementioned size and polarity changes. To define the changes, amino acids were grouped into four polarity groups and three size groups. Polarity groups included non-polar (alanine, glycine, isoleucine, leucine, proline, valine, methionine, phenylalanine), polar neutral (asparagine, glutamine, serine, threonine, tyrosine, cysteine, tryptophan), polar acidic (glutamic acid, aspartic acid), and polar basic (arginine, histidine, lysine). Size groups were defined based on the relative side chain size previously defined by Epstein[35] and included bulky (tryptophan, tyrosine, arginine, phenylalanine), intermediate (histidine, glutamic acid, glutamine, lysine, methionine, asparagine, leucine, isoleucine, proline), and small (cysteine, threonine, valine, alanine, glycine).

ii) Continuous and non-directional (Grantham's distance): A value from 5 (most similar, leucine-isoleucine) to 215 (most dissimilar, cysteine-tryptophan) was assigned to each amino acid substitution mapping it to Grantham's distance matrix. This distance depends on three properties: composition, polarity, and molecular volume; and is independent of the directionality of the change (e.g. leucine > isoleucine is the same as isoleucine > leucine).

iii) Continuous and directional (Epstein's coefficient of difference): A value from 0 (most similar) to 1 (most different) was assigned to each amino acid substitution mapping it to Epstein's coefficient of difference matrix. This coefficient depends on the polarity and size of the replaced amino acids and takes into account directionality (e.g. leucine > tyrosine is 0.28 and tyrosine > leucine is 0.22).

The number of variants and bioactivity data was subsequently calculated per substitution type for different subsets of proteins. For variants with multiple substitutions, each substitution was considered, and therefore accounted for, separately.

Amino acid substitution location analysis

Amino acid substitutions in a protein were defined by their location within the protein with respect to its binding pocket. To this end, each protein was mapped by its UniProt accession code to the available PDB structures with a co-crystalized ligand, which were downloaded as PDB files. Next, for each structure, the structure's first chain with the crystalized ligand was extracted and, for that chain, the ligand's coordinates in the PDB file were retrieved. Based on these coordinates, the ligand's centre of geometry (centroid) was calculated. Similarly, the centroid of each residue in the chain was also calculated. Finally, the distance between the ligand's centroid and each residue's centroid was computed, and the average distance was calculated for each residue across all PDB structures available for a protein. The average distance between the substituted residues' centroid and the ligand's centroid was subsequently used as a metric to differentiate variants based on the location of the substituted residue in the protein. Of note, the average distance between centroids will by definition be larger than the shortest distance to the ligand, which is generally considered when using distances of 5 Å to define the binding pocket. This metric was constructed to be as ligand-agnostic as possible, which in turn leads to non-generalizable distance ranges and should therefore be considered carefully (as an example two ligands with different sizes and binding modes leading to different distances to key residues in EGFR are presented in **Supplementary Figure 15**). In variants with multiple substitutions, each substitution was considered separately. For the analysis of HIV-1 RNaseH-RT (Q72547), only the first of two retrieved PDB codes (2JLE and 3HYF) was used to annotate substitutions located in the reverse transcriptase domain (**Supplementary Figure 16**).

Common subset design

The analysis of variant bioactivity data was done on common subsets of small molecules to ensure fair and accurate comparisons between distributions (**Supplementary Figure 17**). When possible, fully dense common subsets were computed, where all compounds of the subset had been tested on all annotated variants. More typically, non-fully dense common subsets - referred to as common subsets - were defined for each *accession* by first keeping molecules that meet a threshold of being tested on a minimum number of variants. For further analysis, this minimum variant threshold was set to at least two variants. Secondly, variant

coverage was calculated as the percentage of molecules in the subset that were tested on a specific variant. Subsequently, variants above a certain coverage threshold were kept for analysis. Ideally, variant coverage would be set to 100% but, due to high data sparsity, it was set to 20% for analysis.

To increase the density of the common subset, a strategy was introduced where similarity-based filters were used for calculating the minimum variant and the variant coverage thresholds. To obtain these similarity-expanded common subsets, we first computed pair-wise Tanimoto similarities for all molecules in our dataset. Then, we assigned to each molecule a similarity group containing all molecules with a Tanimoto similarity above a certain threshold (0.80). Next, we computed common subset thresholds considering not only true activity points but also activity points in the similarity groups. This is, for threshold calculation a non-existing activity point of molecule X in variant A was counted as existing if compound Y, similar to X, was tested in variant A.

Common subsets were also computed to enable full-panel bioactivity analysis of proteins without a true fully dense common subset. For example, for EGFR (P00533), a bioactivity analysis subset was derived from a common subset computed with a minimum variant threshold of three and a variant coverage of 10%. For HIV-1 RNaseH-RT (Q72574), from a common subset for variants with a compound coverage greater than 3%. For IDHC (O75874), from a common subset for compounds tested on at least two variants and variants with a compound coverage greater than 20%. Finally, for epigenetic regulator BRD4 (O60885), from a common subset for variants with a compound coverage greater than 2%.

The differences between the bioactivity distributions across different types of common subsets were analyzed by calculating the Wasserstein distance between distributions of the *pchembl_value_Mean* variable separately for the WT and all variants combined.

Molecular clustering and visualization

Small molecules in a subset of compounds were clustered using the Butina algorithm to represent their structural similarity across the subset. Starting from compounds represented by canonical SMILES, molecular objects were generated using RDKit. Subsequently, RDKit Daylight-like topological fingerprints were generated and the Tanimoto distance matrix was calculated based on these. Finally, the Butina cluster algorithm was applied to the similarity matrix with a varying cutoff for each subset to minimize the number of single-element clusters. Clusters generated to analyse variant bioactivity distributions in **Figure 7** were computed for subsets including all compounds tested on at least two variants and a Butina cluster cutoff of 0.5. Clusters generated to analyse the full-panel bioactivity differences of compounds in the

EGFR (P00533; **Figure 6**), BRD4 (O60885), and IDHC (O75874) bioactivity analysis subsets were computed for said bioactivity analysis subsets with a Butina cluster cutoff of 0.7. For HIV-1 RNaseH-RT (Q72574), the cluster cutoff was set to 0.5.

To visualize the molecules in a subset of compounds, 2D molecular representations were computed with RDKit. Molecular substructures of interest were matched and highlighted in red. These included either the largest ring system in the molecule or the atoms corresponding to the maximum common substructure of all the compounds in a given cluster.

Variant bioactivity distribution analysis

The distribution of bioactivity values across variants per protein was analysed for three different types of subsets: i) modelling, ii) common, and iii) Butina clusters. These subsets were computed to capture differences in bioactivity across variants covering, respectively, i) all compounds tested on a given protein, ii) a common subset of compounds tested across variants, and iii) different areas of the chemical space tested on a given protein. Common subsets were computed as defined in the section *Common subset design.* In all cases, univariate pChEMBL value distributions were plotted using kernel density estimations in Seaborn for each variant present in the protein subset.

To give an idea of the data sparsity across variants in the different subsets, variant coverage was calculated and reported as defined in the section *Common subset design.* To summarize the bioactivity distribution information, the mean and standard deviation pChEMBL value for each variant was calculated. Moreover, the difference in mean pChEMBL value with respect to the WT was calculated for each variant by subtracting the variant's mean pChEMBL value from the WT's mean pChEMBL value.

Modelling of bioactivities

Three sets were considered for modelling with machine learning. The first set consisted of the original set of bioactivity values obtained for both WT and variant proteins. The second set consisted of data points relating to the WT protein sequences only. Finally, the third set consisted of the similarity-derived common subsets.

All three sets were independently modelled with a quantitative structure-activity relationship (QSAR) model for each accession without any protein sequence-derived feature and with a proteochemometrics (PCM) model for all accessions altogether with sequence features. Protein sequences containing other than the 20 natural amino acids were not considered for modelling with PCM. The collected negative logarithmically scaled bioactivities values were

modelled using the XGBoost (version 1.7.5) implementation of gradient-boosted regression trees[73]. Molecules were represented with the 777 physicochemical and topological Mold2 molecular descriptors[74]. Unaligned protein sequences were described with ProDEC[75] by splitting them into 50 equal parts and averaging the first three principal components (PCs) of Sandberg *et al.*'s amino acid descriptors over each part and over the entire sequence for each of the three PCs, resulting in 153 features (50 parts x 3 PCs + 3 averages PCs)[11,76,77]. Models were 5-fold cross-validated using a random split with a random seed set to 1234 and using a leave-one-out strategy applied for each sequence variant (LOVO). Accessions with less than five data points were disregarded for QSAR modelling and data points related to only one variant were not considered for PCM modelling. Applicability domains were derived using MLChemAD (version 1.2.0) with isolation forests by fitting the training subsets and evaluating them on the Enamine Hit Locator Library (downloaded on 24/01/2024), emulating a typical real-world virtual screening. Finally, the performances of cross-validated models were statistically evaluated between *'PCM-All'*, *'QSAR-WT`*, and *'QSAR-All`* models using Friedman's test for repeated samples using Scipy (version 1.11.2). Significant differences (p-value<0.05) were further investigated using pairwise uncorrected post-hoc Conover-Friedman tests (p-value<0.05) using scikit_posthocs (version 0.8.0).

Data and code availability statement

The data underlying the results and conclusions derived from this work are available online at Zenodo (https://doi.org/10.5281/zenodo.11236694). The Python code and Jupyter Notebooks used to compile and analyse these data is available and maintained on GitHub (https://github.com/CDDLeiden/chembl_variants).

Acknowledgements

The research leading to these results has received funding from a Strategic Award from the Wellcome Trust [104104/A/14/Z], and a Biomedical Resources Grant from the Wellcome Trust [218244/Z/19/Z].

References

1. Leelananda, S. P. & Lindert, S. Computational methods in drug discovery. *Beilstein J. Org. Chem.* **12**, 2694–2718 (2016).
2. Hessler, G. & Baringhaus, K.-H. Artificial Intelligence in Drug Design. *Molecules* **23**, 2520 (2018).
3. Mager, P. P. Theoretical approaches to drug design and biological activity: critical comments to the use of mathematical methods applied to univariate and multivariate quantitative structure-activity relationships (QSAR). *Med Res Rev* **2**, 93–121 (1982).
4. Matsuzaka, Y. & Uesawa, Y. Ensemble Learning, Deep Learning-Based and Molecular Descriptor-Based Quantitative Structure-Activity Relationships. *Molecules* **28**, 2410 (2023).
5. Trapotsi, M.-A. *et al.* Comparison of Chemical Structure and Cell Morphology Information for Multitask Bioactivity Predictions. *J Chem Inf Model* **61**, 1444–1456 (2021).
6. Norinder, U., Spjuth, O. & Svensson, F. Using Predicted Bioactivity Profiles to Improve Predictive Modeling. *J Chem Inf Model* **60**, 2830–2837 (2020).
7. Li, Y. *et al.* Introducing block design in graph neural networks for molecular properties prediction. *Chem Eng J* **414**, 128817 (2021).
8. Bongers, B. J. *et al.* Proteochemometric Modeling Identifies Chemically Diverse Norepinephrine Transporter Inhibitors. *J Chem Inf Model* **63**, 1745–1755 (2023).
9. Bongers, B. J., IJzerman, A. P. & Van Westen, G. J. P. Proteochemometrics – recent developments in bioactivity and selectivity modeling. *Drug Discov. Today Technol.* **32**, 89–98 (2019).
10. Kim, P. T., Winter, R. & Clevert, D.-A. Unsupervised Representation Learning for Proteochemometric Modeling. *Int J Mol Sci* **22**, 12882 (2021).
11. Lenselink, E. B. *et al.* Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. *J. Cheminformatics* **9**, 45 (2017).
12. Zakharov, A. V. *et al.* Novel Consensus Architecture to Improve Performance of Large-Scale Multitask Deep Learning QSAR Models. *J. Chem. Inf. Model.* **59**, 4613–4624 (2019).
13. Cortes-Ciriano, I. *et al.* Proteochemometric modeling in a Bayesian framework. *J. Cheminformatics* **6**, 35 (2014).
14. Wang, D. D., Xie, H. & Yan, H. Proteo-chemometrics interaction fingerprints of protein-ligand complexes predict binding affinity. *Bioinformatics* **37**, 2570–2579 (2021).
15. Sokouti, B. & Hamzeh-Mivehroud, M. 6D-QSAR for predicting biological activity of human aldose reductase inhibitors using quasar receptor surface modeling. *BMC Chem Biol* **17**, 63 (2023).
16. Atas Guvenilir, H. & Doğan, T. How to approach machine learning-based prediction of drug/compound-target interactions. *J Cheminform* **15**, 16 (2023).
17. Mayr, A. *et al.* Large-scale comparison of machine learning methods for drug target prediction on ChEMBL. *Chem Sci* **9**, 5441–5451 (2018).
18. Zhao, L., Wang, W., Sedykh, A. & Zhu, H. Experimental Errors in QSAR Modeling Sets: What We Can Do and What We Cannot Do. *ACS Omega* **2**, 2805–2812 (2017).
19. 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).
20. Tiikkainen, P., Bellis, L., Light, Y. & Franke, L. Estimating error rates in bioactivity databases. *J Chem Inf Model* **53**, 2499–2505 (2013).
21. Kalliokoski, T., Kramer, C., Vulpetti, A. & Gedeck, P. Comparability of mixed IC_{50} data - a statistical analysis. *PLoS One* **8**, e61007 (2013).
22. Kramer, C., Kalliokoski, T., Gedeck, P. & Vulpetti, A. The experimental uncertainty of heterogeneous public K(i) data. *J Med Chem* **55**, 5165–5173 (2012).
23. Geng, C., Vangone, A. & Bonvin, A. M. J. J. Exploring the interplay between experimental methods and the performance of predictors of binding affinity change upon mutations in protein complexes. *Protein Eng Sel* **29**, 291–299 (2016).
24. Feng, C. *et al.* Cancer-Associated Mutations of the Adenosine A2A Receptor Have Diverse Influences on Ligand Binding and Receptor Functions. *Molecules* **27**, 4676 (2022).
25. den Hollander, L. S. *et al.* Impact of cancer-associated mutations in CC chemokine receptor 2 on receptor function and antagonism. *Biochem. Pharmacol.* **208**, 115399 (2023).
26. Hu, Y. *et al.* Naturally occurring mutations of SARS-CoV-2 main protease confer drug resistance to nirmatrelvir. Preprint at *BioRxiv* https://doi.org/10.1101/2022.06.28.497978 (2022).
27. Du, Y. *et al.* Evolution of Multiple Domains of the HIV-1 Envelope Glycoprotein during Coreceptor Switch with CCR5 Antagonist Therapy. *Microbiol Spectr* **10**, e0072522 (2022).

28. Yver, A. Osimertinib (AZD9291)-a science-driven, collaborative approach to rapid drug design and development. *Ann Oncol* **27**, 1165–1170 (2016).
29. Musharrafieh, R., Ma, C. & Wang, J. Discovery of M2 channel blockers targeting the drug-resistant double mutants M2-S31N/L26I and M2-S31N/V27A from the influenza A viruses. *Eur J Pharm Sci* **141**, 105124 (2020).
30. Landrum, G. A. & Riniker, S. Combining IC50 or Ki Values from Different Sources Is a Source of Significant Noise. *J. Chem. Inf. Model.* **64**, 1560–1567 (2024).
31. Leeson, P. D. *et al.* Target-Based Evaluation of "Drug-Like" Properties and Ligand Efficiencies. *J. Med. Chem.* **64**, 7210–7230 (2021).
32. Van Westen, G., Hendriks, A., Wegner, J. K., Ijzerman, A. P. & Van Vlijmen, H. W. T. Significantly Improved HIV Inhibitor Efficacy Prediction Employing Proteochemometric Models Generated From Antivirogram Data. *PLoS Comput Biol* **9**, 1002899 (2013).
33. Christmann-Franck, S. *et al.* Unprecedently Large-Scale Kinase Inhibitor Set Enabling the Accurate Prediction of Compound-Kinase Activities: A Way toward Selective Promiscuity by Design? *J. Chem. Inf. Model.* **56**, 1654–1675 (2016).
34. Béquignon, O. J. M. *et al.* Papyrus: a large-scale curated dataset aimed at bioactivity predictions. *J. Cheminformatics* **15**, 3 (2023).
35. Epstein, C. J. Non-randomness of Ammo-acid Changes in the Evolution of Homologous Proteins. *Nature* **215**, 355–359 (1967).
36. Butina, D. Unsupervised Data Base Clustering Based on Daylight's Fingerprint and Tanimoto Similarity: A Fast and Automated Way To Cluster Small and Large Data Sets. *J. Chem. Inf. Comput. Sci.* **39**, 747–750 (1999).
37. Hunter, F. M. I. *et al.* Drug Safety Data Curation and Modeling in ChEMBL: Boxed Warnings and Withdrawn Drugs. *Chem. Res. Toxicol.* **34**, 385–395 (2021).
38. Bento, A. P. *et al.* An open source chemical structure curation pipeline using RDKit. *J. Cheminformatics* **12**, 51 (2020).
39. Starlinger, J. *et al.* Variant information systems for precision oncology. *BMC Med. Inform. Decis. Mak.* **18**, 107 (2018).
40. Masoudi-Sobhanzadeh, Y., Omidi, Y., Amanlou, M. & Masoudi-Nejad, A. Drug databases and their contributions to drug repurposing. *Genomics* **112**, 1087–1095 (2020).
41. Shafer, R. W. Rationale and Uses of a Public HIV Drug-Resistance Database. *J. Infect. Dis.* **194**, S51–S58 (2006).
42. Sandgren, A. *et al.* Tuberculosis Drug Resistance Mutation Database. *PLoS Med.* **6**, e1000002 (2009).
43. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).
44. Duffy, S. Why are RNA virus mutation rates so damn high? *PLOS Biol.* **16**, e3000003 (2018).
45. Cagan, A. *et al.* Somatic mutation rates scale with lifespan across mammals. *Nature* **604**, 517–524 (2022).
46. Lopez-Bigas, N., De, S. & Teichmann, S. A. Functional protein divergence in the evolution of Homo sapiens. *Genome Biol.* **9**, R33 (2008).
47. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
48. Liu, H., Zhang, B. & Sun, Z. Spectrum of EGFR aberrations and potential clinical implications: insights from integrative pan-cancer analysis. *Cancer Commun.* **40**, 43–59 (2020).
49. Testoni, E. *et al.* Somatically mutated ABL1 is an actionable and essential NSCLC survival gene. *EMBO Mol. Med.* **8**, 105–116 (2016).
50. Śmiech, M., Leszczyński, P., Kono, H., Wardell, C. & Taniguchi, H. Emerging BRAF Mutations in Cancer Progression and Their Possible Effects on Transcriptional Networks. *Genes* **11**, 1342 (2020).
51. Bresler, S. C. *et al.* ALK mutations confer differential oncogenic activation and sensitivity to ALK inhibition therapy in neuroblastoma. *Cancer Cell* **26**, 682 (2014).
52. Artemenko, M., Zhong, S. S. W., To, S. K. Y. & Wong, A. S. T. p70 S6 kinase as a therapeutic target in cancers: More than just an mTOR effector. *Cancer Lett.* **535**, 215593 (2022).
53. Rocha, E. M., Keeney, M. T., Maio, R. D., Miranda, B. R. D. & Greenamyre, J. T. LRRK2 and idiopathic Parkinson's disease. *Trends Neurosci.* **45**, 224–236 (2022).
54. Bracht, J. W. P. *et al.* BRAF Mutations Classes I, II, and III in NSCLC Patients Included in the SLLIP Trial: The Need for a New Pre-Clinical Treatment Rationale. *Cancers* **11**, 1381 (2019).
55. Sunami, T. *et al.* Structural Basis of Human p70 Ribosomal S6 Kinase-1 Regulation by Activation Loop Phosphorylation. *J. Biol. Chem.* **285**, 4587–4594 (2010).
56. Yang, L.-K. & Tao, Y.-X. Alanine Scanning Mutagenesis of the DRYxxI Motif and Intracellular Loop 2 of Human

Melanocortin-4 Receptor. *Int. J. Mol. Sci.* **21**, 7611 (2020).
57. Tate, C. G. & Schertler, G. F. Engineering G protein-coupled receptors to facilitate their structure determination. *Curr. Opin. Struct. Biol.* **19**, 386–395 (2009).
58. Muk, S. *et al.* Machine Learning for Prioritization of Thermostabilizing Mutations for G-Protein Coupled Receptors. *Biophys. J.* **117**, 2228–2239 (2019).
59. Gilmer, T. M. *et al.* Impact of Common Epidermal Growth Factor Receptor and HER2 Variants on Receptor Activity and Inhibition by Lapatinib. *Cancer Res.* **68**, 571–579 (2008).
60. Singh, H., Singh, S., Singla, D., Agarwal, S. M. & Raghava, G. P. S. QSAR based model for discriminating EGFR inhibitors and non-inhibitors using Random forest. *Biol. Direct* **10**, 10 (2015).
61. Burggraaff, L. *et al.* Annotation of allosteric compounds to enhance bioactivity modeling for class A GPCRs. *J. Chem. Inf. Model.* **60**, 4664–4672 (2020).
62. Dera, A. A. *et al.* Identification of Potent Inhibitors Targeting EGFR and HER3 for Effective Treatment of Chemoresistance in Non-Small Cell Lung Cancer. *Molecules* **28**, 4850 (2023).
63. Cortes-Ciriano, I., Murrell, D. S., Van Westen, G. J., Bender, A. & Malliavin, T. E. Prediction of the potency of mammalian cyclooxygenase inhibitors with ensemble proteochemometric modeling. *J. Cheminformatics* **7**, 1 (2015).
64. Hasan, R. & Chu, C. Noise in Datasets: What Are the Impacts on Classification Performance? *Proc. 11th Int. Conf. Pattern Recognit. Appl. Methods* (2022) doi:10.5220/0010782200003122.
65. Kumar, V., De, P., Ojha, P. K., Saha, A. & Roy, K. A Multi-layered Variable Selection Strategy for QSAR Modeling of Butyrylcholinesterase Inhibitors. *Curr. Top. Med. Chem.* **20**, 1601–1627 (2020).
66. Burggraaff, L. *et al.* Successive statistical and structure-based modeling to identify chemically novel kinase inhibitors. *J. Chem. Inf. Model.* **60**, 4283–4295 (2020).
67. Caiafa, C. F., Sun, Z., Tanaka, T., Marti-Puig, P. & Solé-Casals, J. Machine Learning Methods with Noisy, Incomplete or Small Datasets. *Appl. Sci.* **11**, 4132 (2021).
68. Aldeghi, M., Gapsys, V. & De Groot, B. L. Predicting Kinase Inhibitor Resistance: Physics-Based and Data-Driven Approaches. *ACS Cent. Sci.* **5**, 1468–1474 (2019).
69. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618**, 333–341 (2023).
70. An, L. *et al.* Defining the sensitivity landscape of EGFR variants to tyrosine kinase inhibitors. *Transl. Res.* **255**, 14–25 (2023).
71. Burggraaff, L., Van Vlijmen, H. W. T., Ijzerman, A. P. & Van Westen, G. J. P. Quantitative prediction of selectivity between the A1 and A2A adenosine receptors. *J. Cheminformatics* **12**, 1–16 (2020).
72. Andrianov, G. V., Gabriel Ong, W. J., Serebriiskii, I. & Karanicolas, J. Efficient Hit-to-Lead Searching of Kinase Inhibitor Chemical Space via Computational Fragment Merging. *J. Chem. Inf. Model.* **61**, 5967–5987 (2021).
73. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *KDD '16* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016). doi:10.1145/2939672.2939785.
74. Hong, H. *et al.* Mold 2 , Molecular Descriptors from 2D Structures for Chemoinformatics and Toxicoinformatics. doi:10.1021/ci800038f.
75. Béquignon, O. J. M. OlivierBeq/ProDEC: Version 1.0.2. Zenodo https://doi.org/10.5281/ZENODO.7007058 (2022).
76. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): Comparative study of 13 amino acid descriptor sets. *J. Cheminformatics* **5**, 41 (2013).
77. Van Westen, G. J. P. *et al.* Benchmarking of protein descriptor sets in proteochemometric modeling (part 2): Modeling performance of 13 amino acid descriptor sets. *J. Cheminformatics* **5**, 42 (2013).