Chapter 3

Hit Discovery via Unsupervised Learning of Fragment-Protein Complexes

Hit detection is a key step in the early stages of the drug discovery process following the identification of a biological target of interest [26]. A 'hit' compound acts as the starting point for the drug design process where the chemical structure of the hit is progressively optimised towards a candidate drug. Approaches towards hit detection generally involve screening large libraries of compounds, both experimentally and computationally.

One of these methodologies is fragment-based drug design (FBDD). In this approach, very low molecular weight compounds ('fragments' with typically less than 18 nonhydrogen atoms [11]) are screened at high concentrations against the target protein with X-ray crystallography. A fragment screening approach is more likely to deliver hits than screening larger drug-like molecules because low molecular complexity compounds are more likely to possess good complementarity with the target protein [22]. Structures of these fragment-protein complexes can then inspire the design of potent binders, either by expanding a fragment to pick up new intermolecular interactions with active site residues, or merging together different spatially proximal fragments [27, 60]. However, despite showing up in X-ray crystallography, the binding affinity of the fragments themselves is typically low. Therefore, gaining potency by fragment expansion or merging is typically a long journey fraught with false starts.

Recently, advances in X-ray crystallography such as automatic crystal mounting robots, fast detectors, as well as increased accessibility to beamtime are enabling high throughput fragment screens. One can routinely go from screening a small fragment library and detecting a handful of hits, to screening 1000s of fragments with ensembles of 100s of fragments hits spanning the binding site [50, 12]. This substantial increase in data enables a systematic data-driven approach for fragment-based hit discovery.

Our key insight is to reframe fragment-based drug design as signal extraction from noisy data by seeking persistent pharmacophore correlations within a fragment ensemble, rather than looking at individual fragments. This is because a fragment itself has low affinity, thus we need the presence of multiple fragments with the same pharmacophore at a particular region of the binding site to provide statistical confidence.

In this chapter, we employ unsupervised machine learning to learn the spatial distribution of fragment pharmacophores in the binding site. We then use the trained model as a scoring function for virtual screening, picking out molecules with matching pharmacophores. We will first retrospectively validate our model on a dataset of SARS-CoV-2 main protease (Mpro) ligands from COVID Moonshot [56]. We then present prospective results on identifying hits against Mpro and the Mac1 domain of SARS-CoV-2 non-structural protein 3 (nsp3-Mac1) by performing a virtually screen a library of 1.4 billion purchasable compounds from EnamineREAL.

3.1 Unsupervised Learning of Pharmacophore Distributions

To turn fragment hits into a model that predicts whether an unknown ligand will bind potently to the binding site, we employ an interpretation inspired by statistical physics. There are multiple chemical motifs that can engage residues on the binding site. These different modes of engagement can be considered as a statistical distribution. Each interaction between a chemical motif on the fragment and a binding site residue corresponds to an instance of this statistical distribution. We assume that the fragment library broadly covers chemical space, and anticipate that stronger interactions will be sampled and therefore observed more often amongst fragment hits than weaker interactions. Note that an individual fragment is a weak binder – fragment screens are done at a high concentration which forces the equilibrium towards forming fragment-protein complexes enabling detection via crystallography. Therefore, we analyse the statistical distribution of fragment-protein interactions formed by the dense fragment hits, rather than any individual fragment (Figure 3.1a).

To numerically approximate this distribution, we quantify binding interactions by coarse-graining the fragment molecules into hydrogen-bond donor, hydrogen-bond acceptor, and aromatic ring “pharmacophores” (Figure 3.2). These are a simple abstractions of molecular features that can make potent interactions with binding site residues, and is a commonly used tool to interpret the biological activity of ligands [31]. The distribution which we then choose to approximate is the pair-wise distance between these pharmacophores. Computational screening of compounds based on pharmacophore distances is a commonly used technique in medicinal chemistry, though here we are extending this concept to enable a statistical interpretation

of fragment hit. We consider pharmacophore features, rather than specific protein-ligand interactions, so that the downstream model takes the ligand as the input rather than having to perform the additional step of computationally placing the ligand in the binding site.

We utilise kernel density estimation (KDE) [42] to estimate this spatial distribution of pairwise pharmacophore distances. We then score unseen molecules by evaluating pharmacophore

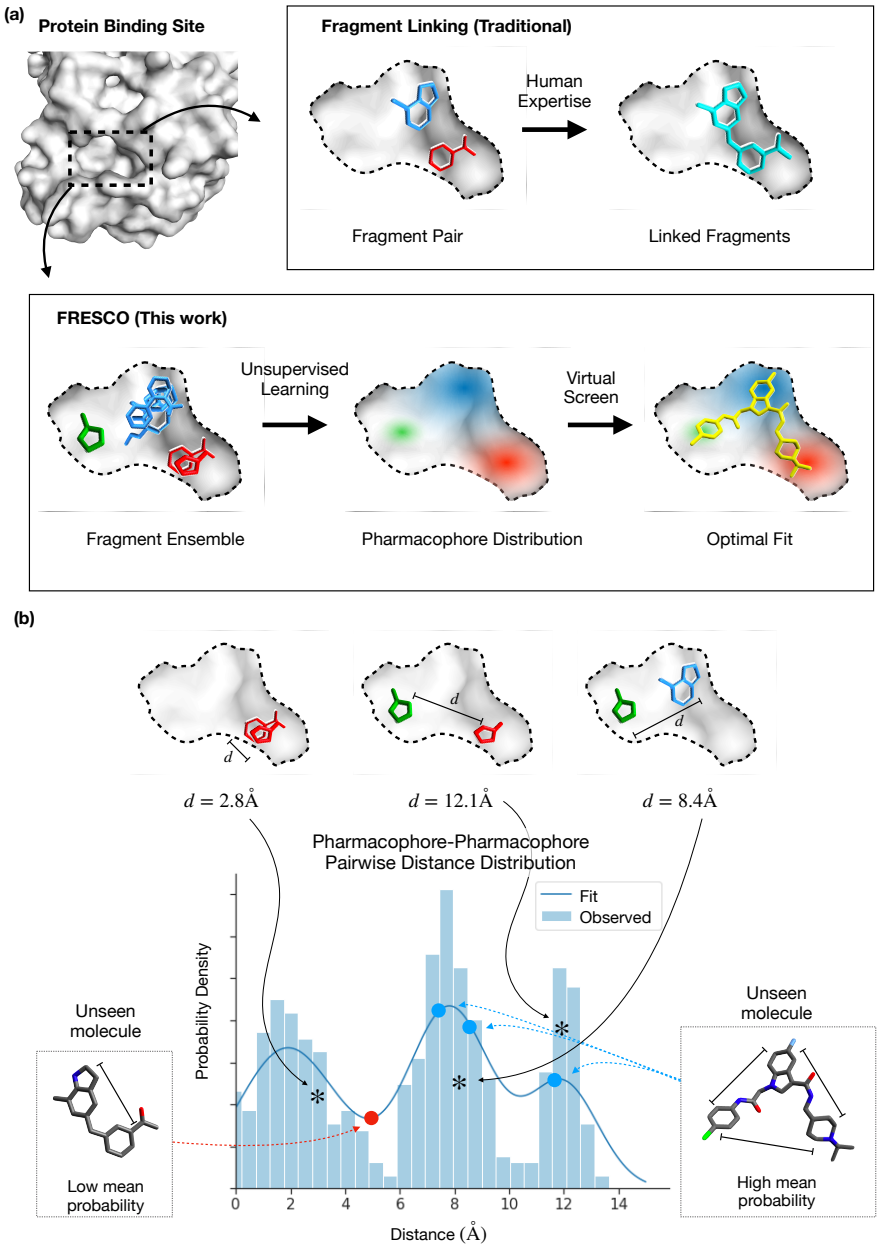

Fig. 3.1 (a) A visual illustration of how FRESCO differs from traditional fragment linking approaches. (b) A visual illustration of how we apply unsupervised learning to fragment ensembles and perform virtual screening of unseen molecules.

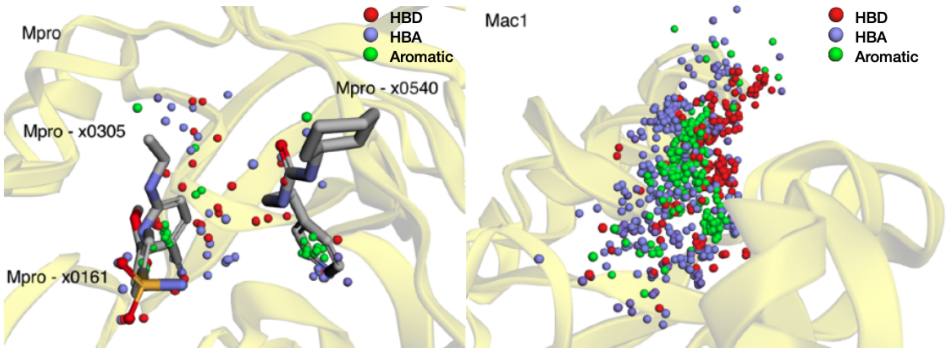

Fig. 3.2 The pharmacophores of the fragment ensemble shown in the 3D binding sites of (a) Mpro, and (b) nsp3-Mac1. The red, blue, and green spheres depict hydrogen bond donors, acceptors, and aromatic pharmacophores respectively. Several of the Mpro fragments are drawn to illustrate the 'origin' of some pharmacophores. None are drawn for nsp3-Mac1 due to the density of pharmacophores in the binding site.

distances within that molecule against the probability distribution of pharmacophore distances derived from the fragment ensemble (Figure 3.1b). We take the mean probability over all of the distances between all possible pharmacophore-pharmacophore pairs as the score for the molecule. This is an unsupervised approach – starting from the results of a crystallographic fragment screen, without any bioactivity data, we can build a model that computationally screens unseen molecules. We term our approach Fragment Ensemble Scoring (FRESCO).

FRESCO conceptually departs from machine learning approaches in the literature for fragment-based hit discovery. These approaches, such as DeLinker [28], SyntaLinker [59], and Develop [29]), as well as data-mining methods such as Fragment Network [21], attempt to grow single fragments or merge only a pair of fragments. They all require expert insights in choosing which fragments to merge, or what pharmacophoric constraints need to obeyed, instead of leveraging all of the information from an ensemble of fragment hits in a data-driven manner.

FRESCO also closes a gap in the burgeoning literature on machine learning for bioactivity prediction [40]. These models cannot be used when no training data exists, as is the case in the hit-finding phase. Thus a new modelling approach – here we employed unsupervised learning – is needed to tackle the "zero-to-one" problem. Although physics-based model of ligand-protein binding such as docking [36, 3, 15] can be used in the absence of any bioactivity data, FRESCO crucially incorporates information from the fragment screen on preferential interactions between regions of the binding site and the fragment pharmacophores.

We validate our approach by performing a retrospective study on historical data, as well as embarking on prospectively campaigns on two different protein targets. Retrospective tests or benchmarks, typically the only method used to compare machine learning models, are

insufficient for measuring the impact of incorporating the model in the decision-making process of compound selection in drug discovery [32]. Thus we go beyond typical model development and undertake a prospective search for hit molecules using only FRESCO to obtain a more realistic measure of its performance.

3.1.1 Model Implementation

To train our model, we first process a set of experimental fragment-protein complexes. In this particular work, we downloaded structures from the `Fragalysis` platform [12]. For Mpro, non-covalent fragments from the XChem fragment screen [12] were used while for Mac1 both XChem and UCSF fragment data were used [52].

We then extract the pharmacophore features from the fragment molecules and their corresponding conformer coordinates. Specifically, we used SMARTS pattern matching following default pharmacophore definitions in `RDKit` to extract pharmacophores from the fragment SMILES. The pharmacophores considered are hydrogen bond donors, hydrogen bond acceptors, and aromatic rings. The corresponding coordinates for each pharmacophore are defined as the average over the atoms in the pharmacophore (eg the position of an aromatic pharmacophore from a benzene ring would be the mean of the coordinates of the 6 carbon atoms in the ring). We then compute the pairwise distance matrix between all possible pharmacophore pairs (eg Donor-Donor & Aromatic-Acceptor) between different fragment molecules.

For some fragments, multiple crystallographic poses are recorded. To account for this, we weigh the contribution of each fragment structure to the overall fragment pharmacophore distribution by $\frac{1}{n}$ where n is the number of conformations recorded for each conformer. In addition, we exclude the counting of correlations between pharmacophores from the same fragment - only correlations between different fragments are measured. This is to avoid spurious intra-fragment correlations that are unrelated to binding to the binding site - strong correlations in pharmacophore distribution between multiple independent fragments are indicative of useful binding interactions and these are what we hope to capture with this methodology.

With the processed 3D pharmacophore distributions, we can then fit a FRESCO model by learning the probability distribution of the pairwise distances using kernel density estimation (KDE). The bandwidth for KDE fitting was chosen for each pariwise distribution using the Improved Sheather-Jones algorithm [5] (implemented in `KDEpy`). KDEs of the systems are then constructed using the chosen bandwidths with `scikit-learn` for technical ease of use in evaluating probabilities. The `scikit-learn` implementation relies on a relatively slow tree-based algorithm that searches over the training datapoints - to increase the computational efficiency of inference for virtual screening, computationally fast approximations of the KDEs are made using the `scipy` `interp1d` function.

With a trained FRESCO model, we can then score unseen molecules by evaluating the probability of the pharmacophore distribution of each molecule. Given a set of input molecular conformers (eg from docking), the same processing workflow is used to obtain the 3D pharmacophore distributions of each molecule, and the probability of the distributions are evaluated using the KDEs. The overall score for the molecule is returned as the mean log-probability over all of the pairwise pharmacophore combinations.

3.2 Computational Retrospective Study

To validate FRESCO, we evaluate how our method compares against the computational approach of docking, as well as the human expertise of medicinal chemists. Specifically, we wish to estimate the extent to which FRESCO could have accelerated hit identification in a fragment-based drug discovery campaign. This requires a dataset that is explicitly exhibiting structure-activity data from the fragment-to-lead phase of a campaign to accurately reflect the degree of structural diversity and distribution of molecular activity. Use of data from an early-stage high-throughput screen would exaggerate the diversity of structures explored, while data from the lead-optimisation phase of a campaign would artificially contain many potent molecules.

For this reason, we choose to study the COVID Moonshot campaign [56] which is targeting the SARS-CoV-2 main protease (Mpro). Mpro is a target of interest for antiviral drug design as inhibition of Mpro inhibits viral replication, as shown by the recent clinical successes of Paxlovid and Ensitrelvir [41, 58]. COVID Moonshot is, to our knowledge, the only openly available dataset of fragment-to-lead drug discovery, driven by a community of medicinal chemists, where every structure and associated activity is disclosed. This unique dataset allows us to perform a time-split analysis, focusing on the fragment-to-lead phase.

The Moonshot activity data for the retrospective study was accessed in Mar 22nd 2021. The IC50 values in that dataset, as well as in the prospective study on Mpro were measured from a fluorescence based enzyme activity assay, the details of which are described below. To narrow down the data to moleucles during the fragment-to-lead stage of the Moonshot campaign, we only selected molecules which were designed before September 1st, 2020, which gave us a dataset of 979 compounds.

In addition, molecular docking studies have also been done extensively on molecules from the Moonshot campaign [39, 48]. For our analysis we utilise the same docking protocols as those reported previously for consistency, the details of which can be found in the methods section.

In the hit identification phase of drug discovery, relatively little is known about what ligand-protein interactions are feasible, thus most proposed molecules are unlikely to be active. A meaningful metric for comparing methods in this regime is the top-N "hit rate", which measures the percentage of the top-N predictions which are active. We expect the curve from plotting the hit rate against N of an informative method to be consistently higher than that of a less informative method. For the Moonshot data we set an IC50 (concentration of inhibitor required to inhibit 50% of protein activity) threshold of 5μM for defining a "hit". This threshold is relatively arbitrary and repeated analysis for both lower and higher IC50 thresholds (1-15μM) show similar results.

The baseline hit rate in the dataset i.e. the percentage of compounds with IC50 $< 5\mu$M , is 6.0%. This represents the hit rate of medicinal chemists using traditional and computational tools at their disposal to design compounds for the Moonshot drug discovery campaign. The hit rate for docking is computed by choosing the top-N molecules with the best score. To calculate the hit rate for FRESCO, we first fit a FRESCO model on 23 publicly reported crystallographic structures of non-covalent fragments bound to the SARS-CoV-2 Mpro protein [12] and score the whole dataset using the fitted FRESCO model.

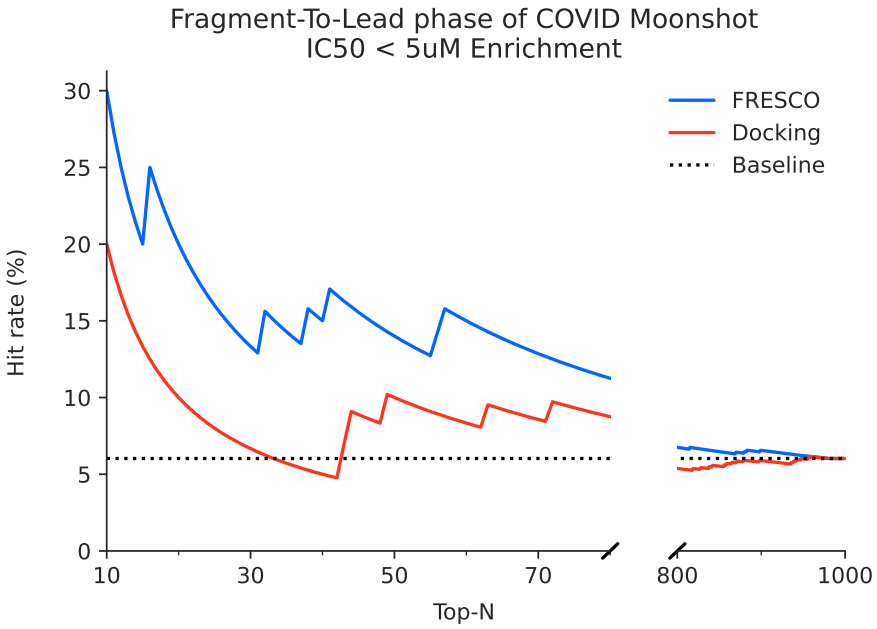

Fig. 3.3 FRESCO is able to retrospectively perform hit detection. High hit rates are achieved relative to docking and the human expert baseline when ranking molecules from the fragment-to-lead phase of COVID Moonshot.

Figure 3.3 shows that FRESCO achieves higher hit rates compared to both computational docking and the medicinal chemists. Looking at the top-5% of the molecules ($N < 50$),

FRESCO has a hit rate of 12-30%, roughly 2-5 times that of the medicinal chemists. Hit rates are also higher than baseline for both lower and higher IC50 thresholds (Figure S3). This shows that it is possible to correlate bioactivity with unsupervised learning of fragment pharmacophore distributions, and that FRESCO could accelerate hit detection in a real-world drug discovery campaign. In this retrospective study, FRESCO is standing on the shoulders of medicinal chemists – it is used to rescore compounds that are designed by chemists. Therefore, we next turn to interrogate the performance of FRESCO in a real-world context when it is used to score a large unbiased library of compounds via a series of prospective studies.

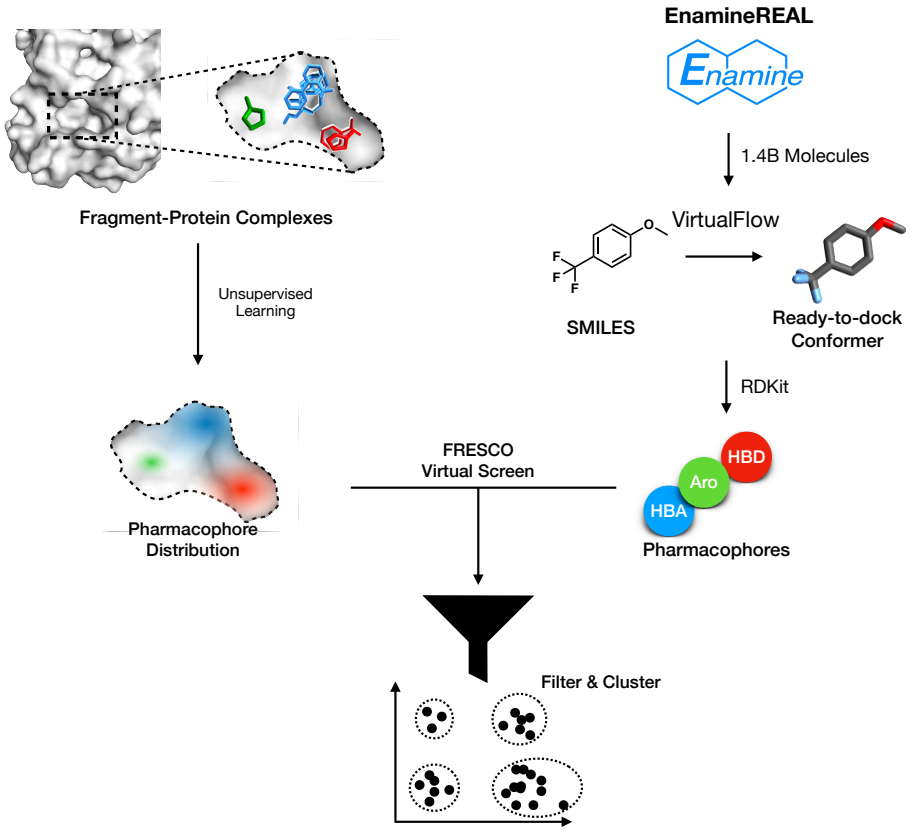

Fig. 3.4 A schematic of the FRESCO screening workflow.

3.3 Prospective hit finding

Building on the results of the retrospective evaluation, we performed a prospective study on Mpro. Rather than rescreening Moonshot compounds, we instead deploy the model to virtually screen a library of commercially available compounds. By synthesising and assaying the

top-ranked compounds, we can evaluate the performance of FRESCO in a real-life use case of hit discovery.

The computational workflow we follow to perform the virtual screening is shown in Figure 3.4. Using a FRESCO model trained on the fragment-protein complexes, we score the library and rank the compounds by score. The top-ranked compounds are then filtered by their physical properties to maximise "drug-likeness", and selected diverse compounds by clustered hit by structural similarity and picking centroids of the most populous clusters.

The library we screen is VirtualFlow, a published dataset of more than 1.4 billion commercially available molecules from EnamineREAL & ZINC15 with pre-generated molecular conformers in a ready-to-dock format [20]. The top-500k predictions were selected and filtered to remove undesirable properties. A series of successive filtering steps were performed: first, only molecules with physical properties in well-understood "lead-like" chemical space [Che] were kept. Secondly, the sum of the number of hydrogen bond donors and hydrogen bond acceptors were constrained to an upper limit of 8. Then, we remove molecules that match known filters for pan-assay interference compounds (PAINS) [4] as well as filters for moieties that are undesirable for medicinal chemistry (eg furan, thiophene, nitro groups). Duplicate tautomers for each molecule are also removed. Finally, for ease of synthetic accessibility, we only consider molecules with less than two chiral centers.

The top-50k molecules remaining from the filtering were then clustered via Butina Clustering [7] with a Tanimoto distance threshold of 0.2. This resulted in 24748 for Mpro. The centroids of the 50 most populous clusters (or the closest purchasable analogue if it wasn't available) were chosen as the candidate compounds. These compounds were ordered for synthesis from Enamine which resulted in 38 successfully made molecules.

Inspecting the cluster centroids favored by FRESCO, we observe typically 2 aromatic moieties connected via an amide or amide isostere. This scaffold is exhibited by three of the initial fragment hits (x0434, x0678, x1093), with the most of the other fragment hits possessing an aromatic group bound at similar locations (Figure **??**). The most promising compound, WIL-UNI-d4749f31-37, has an IC50 of 25.8μM measured via fluorescence assay while the remaining compounds were found to be weak-to-negligible activity.

To validate compound activity, we synthesized 8 close analogues to demonstrate the existence of responsive Structure-Activity Relationship [24, 38] (Figure **??**). 3 of those compounds, which contained modifications to the 2-hydroxyquinoline substructure of WIL-UNI-d4749f31-37, retained relatively high potency of IC50 $< 100\mu$M with one of them (ALP-UNI-ed5cdfd2-1) exhibiting a lower IC50 of 19.4μM . The remaining 5 compounds which perturbed the benzimidazole functional group of WIL-UNI-d4749f31-3 exhibit decreased potency, with only 20-50% inhibition at a concentration of 99.5μM .

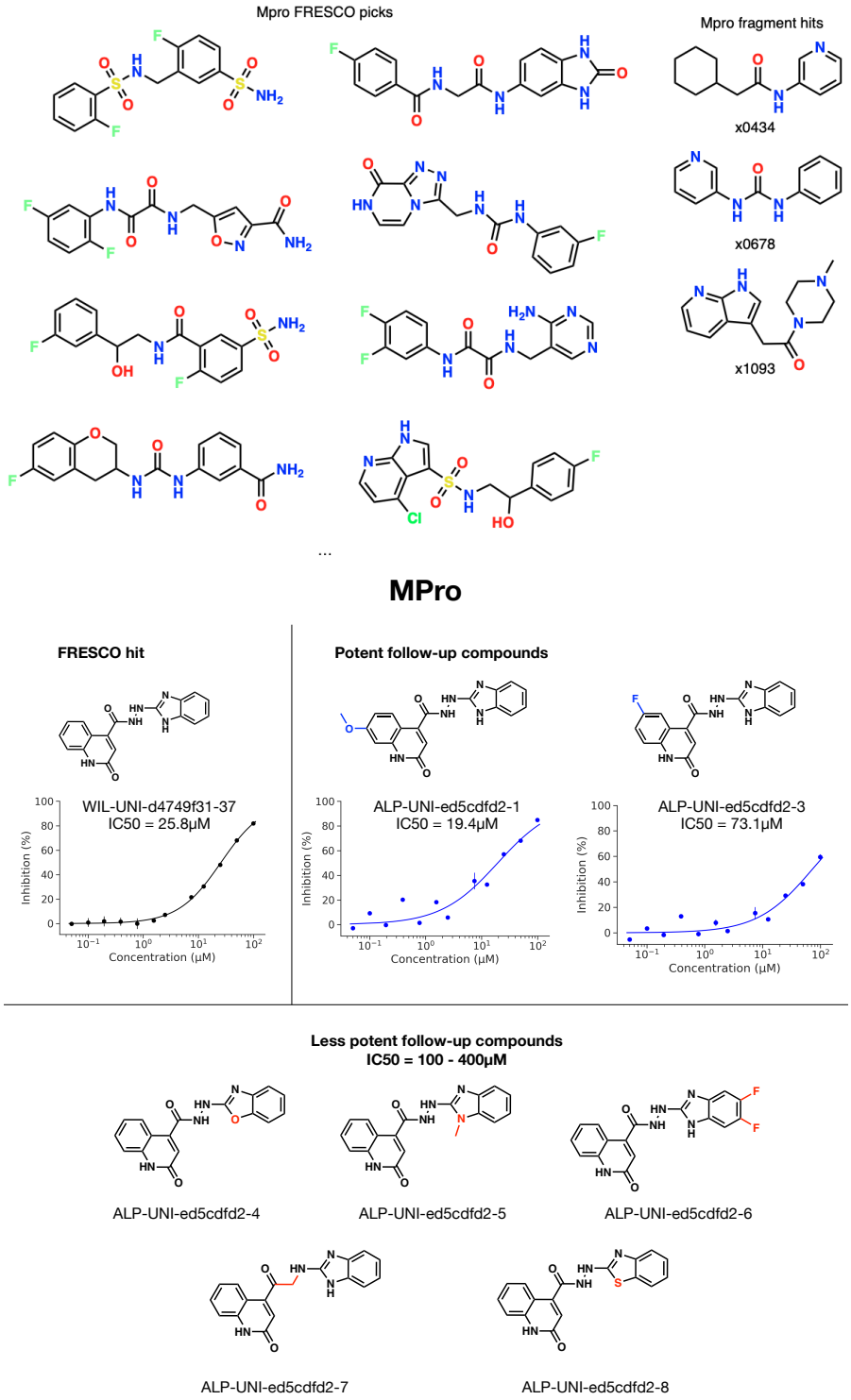

Fig. 3.5 (a) Example structures of cluster centroids after executing the FRESCO screening workflow on Mpro. The molecules favoured by FRESCO tend to have 2 aromatic moieties connected via an amide or an amide isostere, similarly exhibited by three of the initial fragment hits whose structures are also shown. (b) Compound WIL-UNI-d4749f31-37 is identified as a hit against Mpro, with hit confirmation via follow-up compounds demonstrating SAR. Perturbations to the 2-hydroxyquinoline substructure of WIL-UNI-d4749f31-37 led to increased potency while changes to the benzimidazole group consistently decreased potency. Structural differences between the follow-up compounds and WIL-UNI-d4749f31-37 are highlighted in blue/red.

We then turn to SARS-CoV-2 nsp3-Mac1, a structurally unrelated protein target, to demonstrate generalisability of FRESCO in performing hit detection. nsp3-Mac1 is a viral ADP-ribosylhydrolase which counteracts host immune response by cleaving ADP-ribose that is transferred to viral proteins by host ADP-ribosyltransferases. Unlike Mpro, there is no potent chemical matter against nsp3-Mac1. As such, this is a novel first-in-class biological target.

Repeating the FRESCO workflow on a fragment screen against Mac1 [52], we obtained 22358 clusters of top-ranked compounds and successfully made 52 molecules. We find that the molecules favored by FRESCO tend to contain a HBA-HBD pair that is spatially proximal within a hetrocyclic motif. This mimics adenosine, a core in the natural substrate, and this motif is shared in many of the initial fragment hits (Figure 3.6). We successfully ordered and assayed 52 of the compounds identified by FRESCO (see SI for the whole library). Two of the compounds show non-negligible activity at high concentration - at 250μM , compound Z5551425673 (as a racemic mixture) has an inhibition of 30.1%, while compound Z1102995175 has 24.8%.

In addition, an X-ray crystallographic screen was also run on the compounds revealing the structure of Z5551425673 (as the S-stereoisomer) bound to the active site (Figure 3.7). Crystal

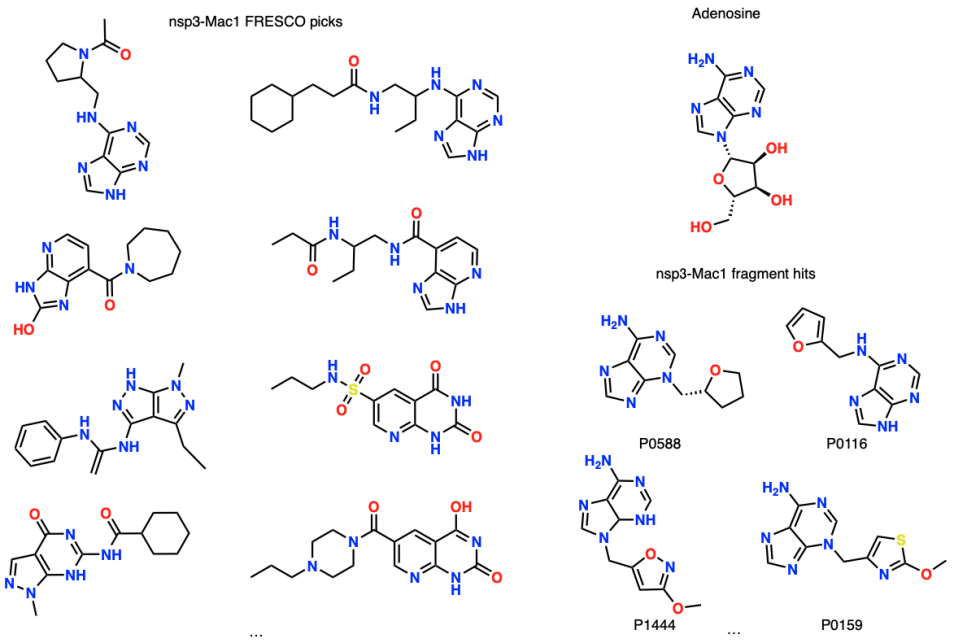

Fig. 3.6 Example structures of cluster centroids after executing the FRESCO screening workflow on nsp3-Mac1. The molecules favoured by FRESCO tend to contain an acceptor-donor pair spatially proximal to a hetrocyclic motif. This mimics adenosine, a core in the natural substrate.This motif is also shared in many of the initial fragment hits, with some example structures shown in the figure.

structures of 9 other compounds chosen via the FRESCO workflow were also obtained though they did not show notable inhibition via HTRF assay. The orthogonal experimental assay and crystal structure results confirm that Z5551425673 is a hit.

As with Mpro, 11 close analogues to Z5551425673 were ordered to explore the structure-activity relationship of the hit and ensure that the compound is not a singleton. 4 compounds perturbing the aliphatic tail substructure had relatively negligible effect while the remaining compounds perturbing the purine group led to a large drop in activity (Figure 3.8). These sets

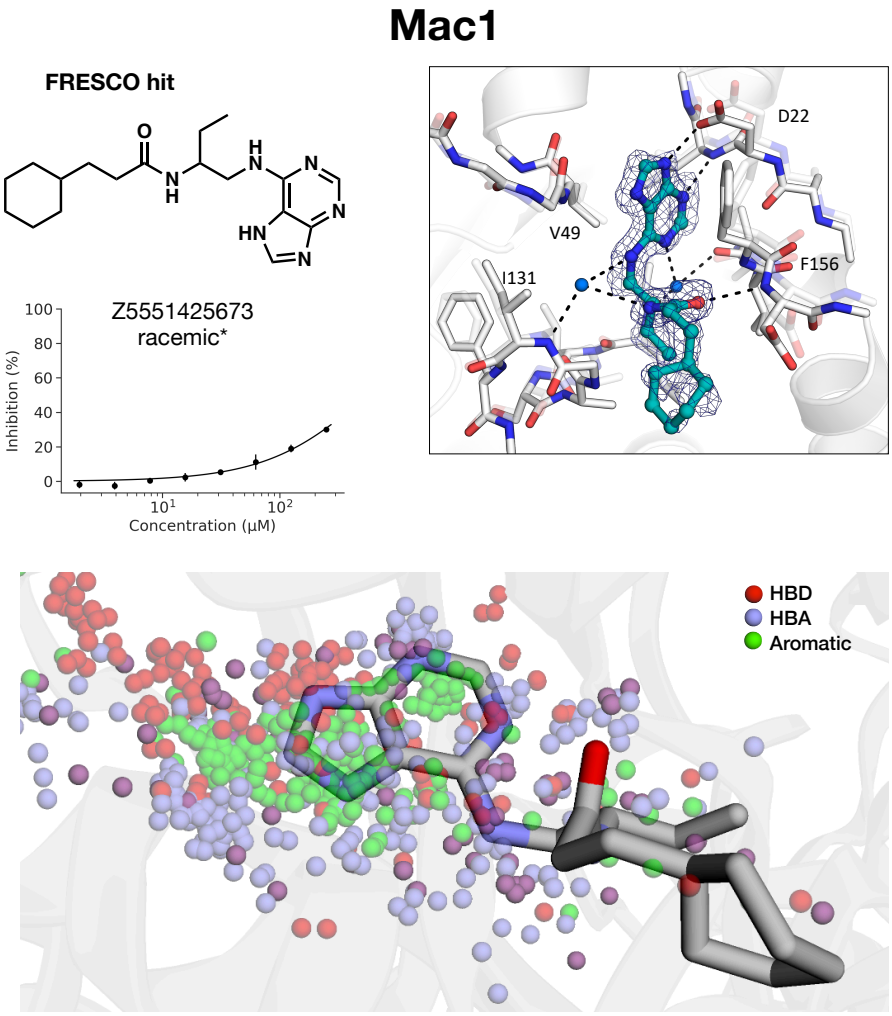

Fig. 3.7 (a) Compound Z5551425673 is identified as a hit against Mac1 via HTRF assay, with (b) hit confirmation via resolution of a crystal structure of Z5551425673 (colored in cyan) bound to the Mac1 active site. (c) The pharmacophores of Z5551425673 match those exhibited by the fragment hits as highlighted by overlaying the bound structure of Z5551425673 (PDB 7FR2) on the distribution of pharmacophores from the fragment ensemble. Note that some functional groups can be regarded as both hydrogen-bond acceptor (blue) and hydrogen-bond donor (red) pharmacophores and hence they are illustrated as purple.

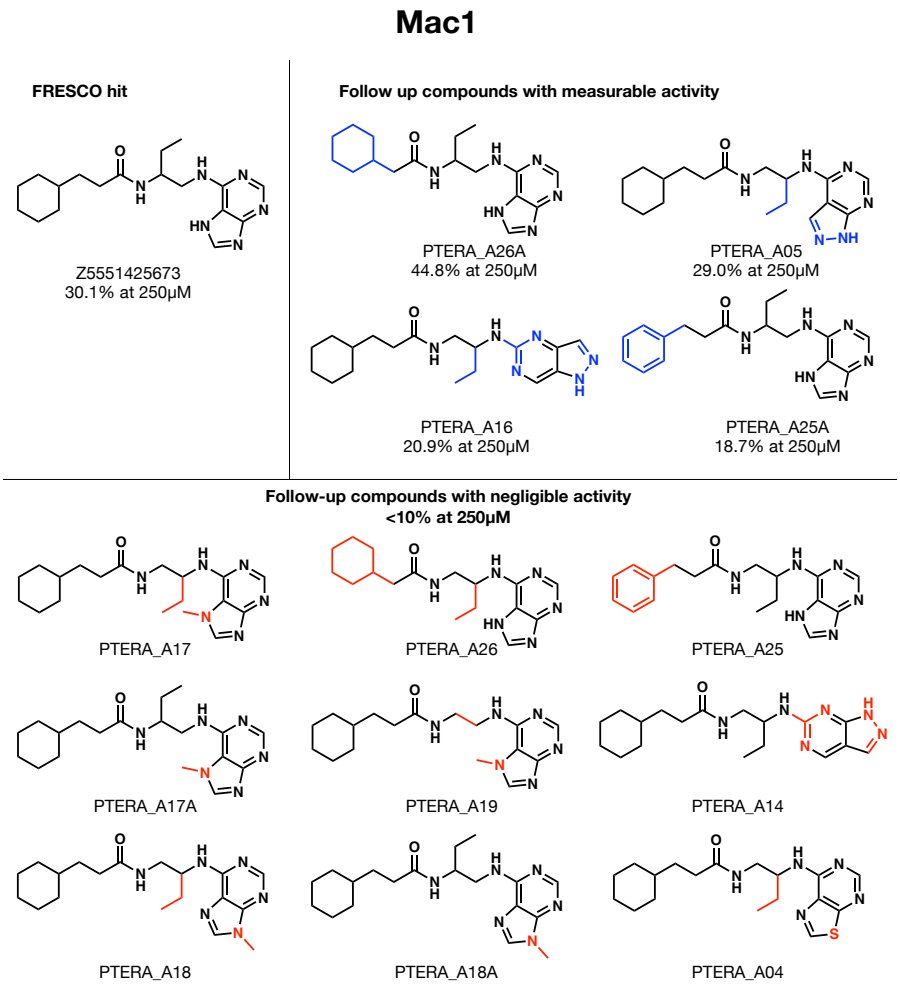

Fig. 3.8 Close analogues around the hit compound identified by FRESCO, Z5551425673, reveals structure-activity relationship which derisks singleton artefacts.

of molecules, still weak in potency, are potentially promising starting points for a hit expansion campaign.

3.4 Discussion

Here we show that the combination of computational statistics with high-throughput structural biology and large libraries of purchasable fragment-like molecules unlocks a powerful tool in hit discovery. Going beyond classical fragment-based drug design, which involves merging or expanding a small set of fragments, we derived a statistical framework that leverages dense fragment hits to build potent inhibitors. Whilst individual fragments are weak binders, our key insight is that a fragment-protein interaction is likely to be significant if there are

multiple fragments making similar interactions. Therefore, by picking out these persistent interactions, we can discern the salient chemical motifs which make favourable interactions with the binding site. Specifically, we coarse-grained fragments into pharmacophores, and infer the distribution of pairwise distances between pharmacophores using Kernel Density Estimation. We then screen large libraries of purchasable compounds against this fragment-derived pharmacophore distribution. We retrospectively validated our method using data from The COVID Moonshot, an open science drug discovery campaign against the SARS-CoV-2 main protease, and prospectively discovered new hits against SARS-CoV-2 main protease and nsp3-Mac1.

More generally, we note that our method does not require the observation of affinity data in order to infer potency. This is done by employing an unsupervised machine learning approach on unlabelled structural biology data. As the throughput of structural biology increases, we hope that an unsupervised approach may unlock novel ways of overcoming data limitations in the protein-ligand affinity prediction problem.

Finally, although prospective studies demonstrated FRESCO's ability to identify hits, we note that the hit rate and potency of the identified hits are both lower than the retrospective experiments. This highlights the importance of prospective validation in machine learning – retrospective studies are biased by the fact that the model is rescoring "reasonable" design from medicinal chemists, whereas in prospective evaluations, the model is used to score the large chemical space without further inductive biases. Future efforts to improve FRESCO should seek to include further inductive biases, for example incorporating physics-based constrains such as docking to filter FRESCO outputs, as well as solidifying a human-in-the-loop approach to select top hits.

References

[Che] Chemspace: Lead-like compounds.

[2] Agarwal, S., Dugar, D., and Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731.

[3] Alon, A., Lyu, J., Braz, J. M., Tummino, T. A., Craik, V., O'Meara, M. J., Webb, C. M., Radchenko, D. S., Moroz, Y. S., Huang, X.-P., Liu, Y., Roth, B. L., Irwin, J. J., Basbaum, A. I., Shoichet, B. K., and Kruse, A. C. (2021). Structures of the

$$\sigma$$

2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764.

[4] Baell, J. B. and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740. PMID: 20131845.

[5] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957.

[6] Brown, N., McKay, B., Gilardoni, F., and Gasteiger, J. (2004). A graph-based genetic algorithm and its application to the multiobjective evolution of median molecules. *Journal of chemical information and computer sciences*, 44(3):1079–1087.

[7] Butina, D. (1999). Unsupervised data base clustering based on daylight's fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750.

[8] Cannalire, R., Cerchia, C., Beccari, A. R., Di Leva, F. S., and Summa, V. (2020). Targeting sars-cov-2 proteases and polymerase for covid-19 treatment: State of the art and future opportunities. *Journal of medicinal chemistry*.

[9] Chodera, J., Lee, A. A., London, N., and von Delft, F. (2020). Crowdsourcing drug discovery for pandemics. *Nature Chemistry*, 12(7):581–581.

[10] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2019). Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie - International Edition*, pages 2–38.

[11] Davis, B. J. and Roughley, S. D. (2017). Chapter eleven - fragment-based lead discovery. In Goodnow, R. A., editor, *Platform Technologies in Drug Discovery and Validation*, volume 50 of *Annual Reports in Medicinal Chemistry*, pages 371–439. Academic Press.

[12] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F., and Walsh, M. A. (2020a). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature Communications*, 11(1):5047.

[13] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., et al. (2020b). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature communications*, 11(1):1–11.

[14] Duffy, N. P. (2010). Molecular property modeling using ranking. US Patent 7,702,467.

[15] Fink, E. A., Xu, J., Hübner, H., Braz, J. M., Seemann, P., Avet, C., Craik, V., Weikert, D., Schmidt, M. F., Webb, C. M., Tolmachova, N. A., Moroz, Y. S., Huang, X.-P., Kalyanaraman, C., Gahbauer, S., Chen, G., Liu, Z., Jacobson, M. P., Irwin, J. J., Bouvier, M., Du, Y., Shoichet, B. K., Basbaum, A. I., and Gmeiner, P. (2022). Structure-based discovery of nonopioid analgesics acting through the

$$\alpha_{2A}$$

-adrenergic receptor. *Science*, 377(6614):eabn7065.

[16] Gao, W. and Coley, C. W. (2020). The synthesizability of molecules proposed by generative models. *Journal of Chemical Information and Modeling*, 60(12):5714–5723.

[17] Gehrtz, P., Marom, S., Bührmann, M., Hardick, J., Kleinbölting, S., Shraga, A., Dubiella, C., Gabizon, R., Wiese, J. N., Müller, M. P., Cohen, G., Babaev, I., Shurrush, K., Avram, L., Resnick, E., Barr, H., Rauh, D., and London, N. (2022). Optimization of covalent mkk7 inhibitors via crude nanomole-scale libraries. *Journal of Medicinal Chemistry*, 65(15):10341–10356.

[18] Gironda-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. (2021). Dna-encoded chemical libraries: A comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279.

[19] Gómez-Bombarelli, R., Wei, J. N., Duvenaud, D., Hernández-Lobato, J. M., Sánchez-Lengeling, B., Sheberla, D., Aguilera-Iparraguirre, J., Hirzel, T. D., Adams, R. P., and Aspuru-Guzik, A. (2018). Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276.

[20] Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., and Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668.

[21] Hall, R. J., Murray, C. W., and Verdonk, M. L. (2017). The fragment network: A chemistry recommendation engine built using a graph database. *Journal of Medicinal Chemistry*, 60(14):6440–6450.

[22] Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 41(3):856–864.

[23] Hartenfeller, M., Zettl, H., Walter, M., Rupp, M., Reisen, F., Proschak, E., Weggen, S., Stark, H., and Schneider, G. (2012). Dogs: reaction-driven de novo design of bioactive compounds. *PLoS Comput Biol*, 8(2):e1002380.

[24] Hermann, J. C., Chen, Y., Wartchow, C., Menke, J., Gao, L., Gleason, S. K., Haynes, N.-E., Scott, N., Petersen, A., Gabriel, S., Vu, B., George, K. M., Narayanan, A., Li, S. H., Qian, H., Beatini, N., Niu, L., and Gan, Q.-F. (2013). Metal impurities cause false positives in high-throughput screening campaigns. *ACS Medicinal Chemistry Letters*, 4(2):197–200.

[25] Howard, J. et al. (2018). fastai. https://github.com/fastai/fastai.

[26] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249.

[27] Ichihara, O., Barker, J., Law, R. J., and Whittaker, M. (2011). Compound design by fragment-linking. *Molecular Informatics*, 30(4):298–306.

[28] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2020). Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995.

[29] Imrie, F., Hadfield, T. E., Bradley, A. R., and Deane, C. M. (2021). Deep generative design with 3d pharmacophoric constraints. *Chem. Sci.*, 12:14577–14589.

[30] Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., et al. (2020). Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293.

[31] Kaserer, T., Beck, K. R., Akram, M., Odermatt, A., and Schuster, D. (2015). Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20(12):22799–22832.

[32] Kearnes, S. (2021). Pursuing a prospective perspective. *Trends in Chemistry*, 3(2):77–79.

[Landrum] Landrum, G. RDKit: Open-source cheminformatics. http://www.rdkit.org.

[34] Lee, A. A., Yang, Q., Sresht, V., Bolgar, P., Hou, X., Klug-McLeod, J. L., Butler, C. R., et al. (2019). Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 55(81):12152–12155.

[35] Liu, Y., Liang, C., Xin, L., Ren, X., Tian, L., Ju, X., Li, H., Wang, Y., Zhao, Q., Liu, H., et al. (2020). The development of coronavirus 3c-like protease (3clpro) inhibitors from 2010 to 2020. *European journal of medicinal chemistry*, page 112711.

[36] Lyu, J., Wang, S., Balius, T. E., Singh, I., Levit, A., Moroz, Y. S., O'Meara, M. J., Che, T., Algaa, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229.

[37] McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., Cuozzo, J. W., Guié, M.-A., Guilinger, J. P., Huguet, C., Hupp, C. D., Keefe, A. D., Mulhern, C. J., Zhang, Y., and Riley, P. (2020). Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866.

[38] Morreale, F. E., Testa, A., Chaugule, V. K., Bortoluzzi, A., Ciulli, A., and Walden, H. (2017). Mind the metal: A fragment library-derived zinc impurity binds the e2 ubiquitin-conjugating enzyme ube2t and induces structural rearrangements. *Journal of Medicinal Chemistry*, 60(19):8183–8191.

[39] Morris, A., McCorkindale, W., Consortium, T. C. M., Drayman, N., Chodera, J. D., Tay, S., London, N., and Lee, A. A. (2021). Discovery of sars-cov-2 main protease inhibitors using a synthesis-directed de novo design model. *Chem. Commun.*, 57:5909–5912.

[40] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., et al. (2020). Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564.

[41] Owen, D. R., Allerton, C. M. N., Anderson, A. S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R. D., Carlo, A., Coffman, K. J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K. S., Gibson, S. A., Greasley, S. E., Hurst, B. L., Kadar, E. P., Kalgutkar, A. S., Lee, J. C., Lee, J., Liu, W., Mason, S. W., Noell, S., Novak, J. J., Obach, R. S., Ogilvie, K., Patel, N. C., Pettersson, M., Rai, D. K., Reese, M. R., Sammons, M. F., Sathish, J. G., Singh, R. S. P., Steppan, C. M., Stewart, A. E., Tuttle, J. B., Updyke, L., Verhoest, P. R., Wei, L., Yang, Q., and Zhu, Y. (2021). An oral sars-cov-2 m^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.

[42] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.

[43] Patel, H., Bodkin, M. J., Chen, B., and Gillet, V. J. (2009). Knowledge-based approach to de novo design using reaction vectors. *Journal of chemical information and modeling*, 49(5):1163–1184.

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[45] Perera, D., Tucker, J. W., Brahmbhatt, S., Helal, C. J., Chong, A., Farrell, W., Richardson, P., and Sach, N. W. (2018). A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434.

[46] Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., and Jung, S.-H. (2016). An overview of severe acute respiratory syndrome–coronavirus (sars-cov) 3cl protease inhibitors: peptidomimetics and small molecule chemotherapy. *Journal of medicinal chemistry*, 59(14):6595–6628.

[PostEra Inc.] PostEra Inc. COVID moonshot. https://postera.ai/covid.

[48] Saar, K. L., Fearon, D., Consortium, T. C. M., von Delft, F., Chodera, J. D., and Lee, A. A. (2021). Turning high-throughput structural biology into predictive inhibitor design. *bioRxiv*.

[49] Santanilla, A. B., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L.-C., Schneeweis, J., Berritt, S., Shi, Z.-C., Nantermet, P., Liu, Y., Helmy, R., Welch, C. J., Vachal, P., Davies, I. W., Cernak, T., and Dreher, S. D. (2015). Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53.

[50] Schiebel, J., Krimmer, S. G., Röwer, K., Knörlein, A., Wang, X., Park, A. Y., Stieler, M., Ehrmann, F. R., Fu, K., Radeva, N., et al. (2016). High-throughput crystallography: reliable and efficient identification of fragment hits. *Structure*, 24(8):1398–1409.

[51] Schneider, P. and Schneider, G. (2016). De novo design at the edge of chaos: Miniperspective. *Journal of medicinal chemistry*, 59(9):4077–4086.

[52] Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R. E., Young, I. D., Martins, L. C., Smith, D. H., Schulze-Gahmen, U., Owens, T. W., Deshpande, I., Merz, G. E., Thwin, A. C., Biel, J. T., Peters, J. K., Moritz, M., Herrera, N., Kratochvil, H. T., null null, Aimon, A., Bennett, J. M., Neto, J. B., Cohen, A. E., Dias, A., Douangamath, A., Dunnett, L., Fedorov, O., Ferla, M. P., Fuchs, M. R., Gorrie-Stone, T. J., Holton, J. M., Johnson, M. G., Krojer, T., Meigs, G., Powell, A. J., Rack, J. G. M., Rangel, V. L., Russi, S., Skyner, R. E., Smith, C. A., Soares, A. S., Wierman, J. L., Zhu, K., O'Brien, P., Jura, N., Ashworth, A., Irwin, J. J., Thompson, M. C., Gestwicki, J. E., von Delft, F., Shoichet, B. K., Fraser, J. S., and Ahel, I. (2021). Fragment binding to the nsp3 macrodomain of sars-cov-2 identified through crystallographic screening and computational docking. *Science Advances*, 7(16):eabf8711.

[53] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. (2019). Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS central science*, 5(9):1572–1583.

[54] Segler, M. H., Kogej, T., Tyrchan, C., and Waller, M. P. (2018). Generating focused molecule libraries for drug discovery with recurrent neural networks. *ACS central science*, 4(1):120–131.

[55] The COVID Moonshot Consortium (2020). Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *bioRxiv, doi:10.1101/2020.10.29.339317*.

[56] The COVID Moonshot Consortium, Achdout, H., Aimon, A., Bar-David, E., Barr, H., Ben-Shmuel, A., Bennett, J., Bilenko, V. A., Bilenko, V. A., Boby, M. L., Borden, B., Bowman, G. R., Brun, J., BVNBS, S., Calmiano, M., Carbery, A., Carney, D., Cattermole, E., Chang, E., Chernyshenko, E., Chodera, J. D., Clyde, A., Coffland, J. E., Cohen, G., Cole,

J., Contini, A., Cox, L., Cvitkovic, M., Dias, A., Donckers, K., Dotson, D. L., Douangamath, A., Duberstein, S., Dudgeon, T., Dunnett, L., Eastman, P. K., Erez, N., Eyermann, C. J., Fairhead, M., Fate, G., Fearon, D., Fedorov, O., Ferla, M., Fernandes, R. S., Ferrins, L., Foster, R., Foster, H., Gabizon, R., Garcia-Sastre, A., Gawriljuk, V. O., Gehrtz, P., Gileadi, C., Giroud, C., Glass, W. G., Glen, R., Glinert, I., Godoy, A. S., Gorichko, M., Gorrie-Stone, T., Griffen, E. J., Hart, S. H., Heer, J., Henry, M., Hill, M., Horrell, S., Huliak, V. D., Hurley, M. F., Israely, T., Jajack, A., Jansen, J., Jnoff, E., Jochmans, D., John, T., Jonghe, S. D., Kantsadi, A. L., Kenny, P. W., Kiappes, J. L., Kinakh, S. O., Koekemoer, L., Kovar, B., Krojer, T., Lee, A., Lefker, B. A., Levy, H., Logvinenko, I. G., London, N., Lukacik, P., Macdonald, H. B., MacLean, B., Malla, T. R., Matviiuk, T., McCorkindale, W., McGovern, B. L., Melamed, S., Melnykov, K. P., Michurin, O., Mikolajek, H., Milne, B. F., Morris, A., Morris, G. M., Morwitzer, M. J., Moustakas, D., Nakamura, A. M., Neto, J. B., Neyts, J., Nguyen, L., Noske, G. D., Oleinikovas, V., Oliva, G., Overheul, G. J., Owen, D., Pai, R., Pan, J., Paran, N., Perry, B., Pingle, M., Pinjari, J., Politi, B., Powell, A., Psenak, V., Puni, R., Rangel, V. L., Reddi, R. N., Reid, S. P., Resnick, E., Ripka, E. G., Robinson, M. C., Robinson, R. P., Rodriguez-Guerra, J., Rosales, R., Rufa, D., Saar, K., Saikatendu, K. S., Schofield, C., Shafeev, M., Shaikh, A., Shi, J., Shurrush, K., Singh, S., Sittner, A., Skyner, R., Smalley, A., Smeets, B., Smilova, M. D., Solmesky, L. J., Spencer, J., Strain-Damerell, C., Swamy, V., Tamir, H., Tennant, R., Thompson, W., Thompson, A., Tomasio, S., Tsurupa, I. S., Tumber, A., Vakonakis, I., van Rij, R. P., Vangeel, L., Varghese, F. S., Vaschetto, M., Vitner, E. B., Voelz, V., Volkamer, A., von Delft, F., von Delft, A., Walsh, M., Ward, W., Weatherall, C., Weiss, S., White, K. M., Wild, C. F., Wittmann, M., Wright, N., Yahalom-Ronen, Y., Zaidmann, D., Zidane, H., and Zitzmann, N. (2022). Open science discovery of oral non-covalent sars-cov-2 main protease inhibitor therapeutics. *bioRxiv*.

[57] Ullrich, S. and Nitsche, C. (2020). The sars-cov-2 main protease as drug target. *Bioorganic & Medicinal Chemistry Letters*, page 127377.

[58] Unoh, Y., Uehara, S., Nakahara, K., Nobori, H., Yamatsu, Y., Yamamoto, S., Maruyama, Y., Taoda, Y., Kasamatsu, K., Suto, T., et al. (2022). Discovery of s-217622, a noncovalent oral sars-cov-2 3cl protease inhibitor clinical candidate for treating covid-19. *Journal of Medicinal Chemistry*, 65(9):6499–6512.

[59] Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. (2020). Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322.

[60] Yu, H. S., Modugula, K., Ichihara, O., Kramschuster, K., Keng, S., Abel, R., and Wang, L. (2021). General theory of fragment linking in molecular design: Why fragment linking rarely succeeds and how to improve outcomes. *Journal of Chemical Theory and Computation*, 17(1):450–462. PMID: 33372778.