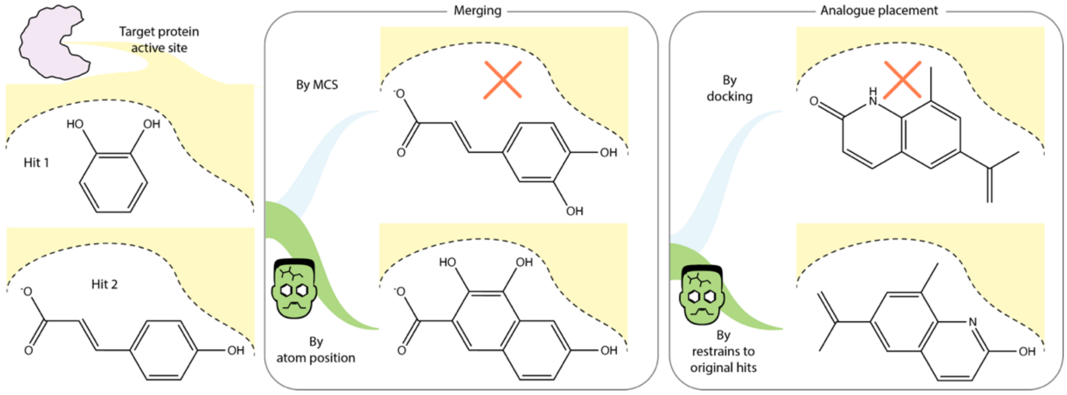
ARTICLE

Fragmenstein: predicting protein-ligand structures of follow-up compounds from known crystallographic fragment hits is more successful when using a strict conserved-binding–based methodology rather than traditional docking-based approaches

Matteo P. Ferla,*[abcd] Rubén Sánchez-García [a], Rachael E. Skyner [ef], Stefan Gahbauer [g], Brian D. Marsden [bd], Jenny C. Taylor [cd], Charlotte M. Deane[a], and Frank von Delft [bdh]

Current strategies centred on either merging or linking initial hits from fragment-based drug design (FBDD) crystallographic screens ignore 3D structural information. We show that an algorithmic approach (Fragmenstein) that 'stitches' the ligand atoms from this structural information together can provide more accurate and reliable predictions for protein-ligand complex conformation than existing methods such as pharmacophore-constrained docking. This approach works under the assumption of conserved binding: when a larger molecule is designed containing the initial fragment hit, the common substructure between the two will adopt the same binding mode. Fragmenstein either takes initial the coordinates of ligands from a experimental fragment screen and stitches the atoms together to produce a novel merged compound, or uses them to predict the complex for a provided compound. The compound is then energy minimised under strong constraints to obtain a structurally correct compound. This method is successful in showing the importance of using the coordinates of known binders when predicting the conformation of follow-ups through a retrospective analysis of the COVID Moonshot data. It has also had a real-world application in hit-to-lead screening, yielding a sub-micromolar merger from inspiration hits in a single round.



1. Introduction

1.1 Limited usage

Traditional methods for linking and merging strategies in FBDD disregard the 3D protein-ligand conformation of promising hit molecules. Fragment-based drug discovery (FBDD) uses small molecules (<250 Da) under the assumption that the information from multiple small molecules is more informative than the information from a low number of larger molecules (typically used in traditional high-throughput screening) in the early hit-to-lead part of drug discovery[1]. This is because small molecules are more likely to have a greater proportion of their potential interaction vectors associating with the protein than the proportion in large molecules, where significant functional parts of the molecule may not interact with the protein at all. Based on this assumption, it should be possible, as part of the FBDD lead-design process, to take the vast amount of protein-ligand interaction information from these smaller proximal molecules to design larger follow-up molecules. This should result in the more efficient design of molecules which possess better binding affinity at a lower cost than lead optimization through structure–activity relationship (SAR) exploration of larger initial hits. It can be argued that this position-based strategy is not currently being exploited to the fullest both when: (i) designing follow-up compounds and (ii) when predicting the binding position of follow-up compounds.

Regardless of whether informative structural information is available for initial hits, by far the most common strategy is to first enumerate follow-up compounds independently of structure, often through similarity or substructure searching, and afterwards employ docking as a conformational filter [2]. As discussed below, the shortcomings of these approaches negatively affect successfulness of the searches.

1.2 Placement

Docking approaches as conformational filters do not sufficiently leverage information from existing protein-ligand structures when predicting the conformation of follow-up compounds. A common method to assess the binding of a candidate molecule is docking. Docking protocols consist of a search algorithm that performs thousands of heuristic iterations assessed by a score function to find the lowest energy predicted position, orientation, and conformation of the ligand in the context of the target protein [3]. Docking protocols find the energetic minimum according to the parameters of the force-field used, but may result in a local energy-minimum conformation that deviates from the one found in the experimental structure. This can occur for a variety of reasons ranging from insufficient or inaccessible sampling of either the ligand or protein conformations to inaccuracies of the physics in the empirical models. A common benchmark to assess the quality of a docking protocol is to "redock" the ligand from an X-ray crystal structure; namely removing the ligand and docking it and comparing the RMSD between the original and the docked ligand. With this approach, even the best algorithms are able to reproduce only roughly half of all compounds docked to an RMSD of less than 2 Å [4]. A method to improve upon this is by constraining the atomic positions to pharmacophores or to key protein-ligand interaction sites, as identified by methods such as hotspot mapping [5]. Most docking algorithms generate a set of protein-independent small molecule conformers before docking which, especially for larger and more flexible small molecules, may greatly diverge from the actual crystallographic protein-bound conformer. Embedding the conformation of two or more FBDD hits within the follow-up compound is straightforward in principle. However, there is currently no approach able to address the issues associated with imperfect overlaps between hits.

1.3 Combination

Combinatorial approaches either are disregard the position of hits or are unable to operate with overlapping hits. When ligands are designed starting from hits (rather than docking a subset of compounds in a dataset), the protein-ligand complex data available from initial hit structures are often still not utilised until after initial enumeration. Such approaches are usually synthon-based, where molecules are broken down into components and then new molecules are designed by combination of components from multiple input ligands. Examples include BRICS decomposition [6] and AutoGrow4 [7]. Neither of these methods consider any 3D structural information from the protein or ligand in the initial enumeration step. DeLinker [8] is an example of a method which takes advantage of the 3D structural information of known ligands by identifying connection vectors between ligands and generating molecules that will fit into that 3D ligand space. However, it is still practically unaware of the protein environment around the ligands it is designing from.

Some methods do consider some spatial information from the protein. For example, GANDI takes protein coordinates into consideration to filter out potential clashes [9], whilst designing linkers in a similar manner to DeLinker. DEVELOP takes this a step further by encoding both protein and ligand conformation into both connectivity (via a graph neural network) and coordinate information (through a voxel occupancy map) in its training to encode pharmacophoric features that can be used to predict new molecules for a protein target not in its training dataset [10]. STRIFE improves upon the predictions made by DEVELOP by also performing docking constrained to hotspot maps to better assess the products after a coarse-grain and a fine-grain step [11]. Although addition of a constrained docking step can greatly improve results over simple ligand-based enumeration and traditional docking, we have found that more stringent constraints are needed to produce reliable results.

None of the methods discussed thus far are able to consider the 3D conformation of overlapping hits. An algorithm that stands out in this respect is BREED [12], implemented within Maestro in the Schrödinger suite, this algorithm joins fragments hits by hybridizing upon spatially overlapping bonds, thus obeying the conformation of the hits. However, it suffers from a few limitations, such as yielding frequently a very limited number of mergers, usage cost, and independence of the protein neighbourhood. Consequently, fragment merging is most often done by eye [13].

2. Methods

2.1 Implementation and algorithm

Fragmenstein is a Python package with few dependencies. The Fragmenstein codebase is a modular Python package that is dependent on RDKit [14] for compound manipulation, PyRosetta [15] for minimisation and some additional open-source purpose-written packages described in the GitHub repository. Its usage does not require external system calls, including the ligand parameterisation for Rosetta, which was rewritten to be both open source and usable within Python 3.6–3.11. Thanks to the limited number of external dependencies, it can be easily deployed in both Linux and MacOS architectures.

Fragmenstein is open source. The open-source codebase for Fragmenstein can be found at https://github.com/matteoferla/Fragmenstein.

Code and data for benchmarks (*vide infra*) available at https://github.com/matteoferla/Fragmenstein-manuscript-data.

Fragmenstein merges or places compounds by stitching together the atoms of the hits. [1415] Fragmenstein has two main functionalities (Figure 1): fragment hit combination and follow-up placement, both constrained by the fragment hits that inspired them. Both these operations require two steps: (i) the creation of a potentially distorted compound whose atoms overlap the inspiration hits and (ii) the energy minimisation of the compound under strong constraints.

The two functionalities can be used as a single continuous workflow as found in the demonstrative notebooks runnable on Google Colab (*viz.* GitHub repository). First fragments are combined (merged/linked) with Fragmenstein, then a search is conducted via the SmallWorld server (https://sw.docking.org/) [16] for purchasable analogues in Enamine REAL or equivalent supplier, and lastly candidate compounds are placed with Fragmenstein in order to be ranked.

The operations performed are described in the GitHub repository. Briefly for both operations mapping of the atoms in different hits is done based on position where each atom can only map to a single atom in the other hit within 2 Å and that for mergers (Figure 1A inset, SI Figure 1), and the last step in both operations is a local and heavily constrained minimisation into the protein pose. For mergers, the rings of the hits are collapsed into a placeholder thus preventing impossible structures (Figure 1A) then the combined atoms are stitched together, rings restored, and any invalid chemistry is corrected. The effect of the operations in the merger can be observed in a test where two rings are placed at specific different distances, yielding different compounds (Figure 1B). For placements, a series of iterative maximum common substructure searches of the desired compound with each hit is used where less stringent searches are constrained by more stringent ones (SI Figure 2) and mapping from searches against the other hits along with certain safeguards constraining bond lengths (Figure 1C).

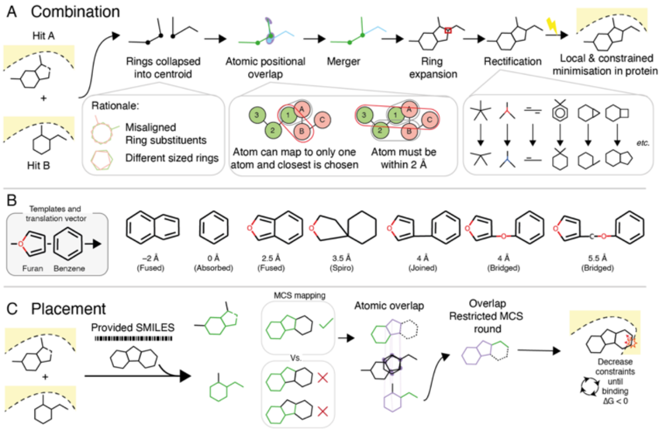
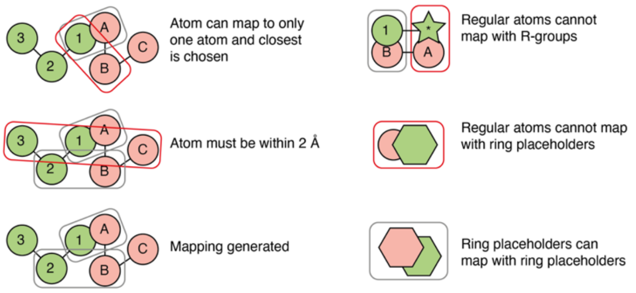
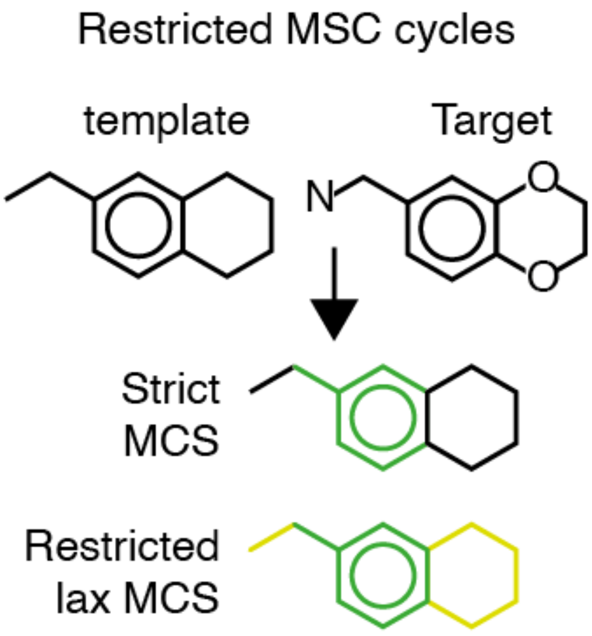


Figure 1. Combination and placement operations and their rules within Fragmenstein. A. Steps in a combination operation. For combinations, the positional overlap is calculated with any ring collapsed. This is done to prevent overlap issues (first inset). Both rules share the atomic positional overlap mapping (middle inset, further detail in Supplementary figure 1). After which, the merger is rectified based on certain rules listed in its GitHub repository. B. The effect of adherence to atomic positions can be seen in a test where a furan and a bezene with centres of mass at different distances yield different molecules ranging from a single ring to two linked rings. C. Steps in a placement operation. The provided compound is mapped to each hit with a multistep MCS scheme (Supplementary figure 2), the mapping with the larger coverage is chosen and the other hits are mapped via a MCS restricted by their atomic overlap with the primary hit. For both combination and positioning, after the stritched together conformer is created it, it is energy minimised locally, with strong constraints and with a topology parametered from an ideal conformer.



Atom can map to only one atom and closest is chosen
Regular atoms cannot map with R-groups
Atom must be within 2 Å
Regular atoms cannot map with ring placeholders
Mapping generated
Ring placeholders can map with ring placeholders

Supplementary Figure 1. Detailed rules employed in determined atomic overlap between two compounds. Atoms cannot map to multiple atoms and cannot map spanning a distance greater than 2 Å. R-groups (dummy atoms) marking covalent attachment points can only map to other R-group. Likewise ring placeholders cannot map to regular atoms.



Supplementary Figure 2. Detailed mapping schema used in the placement operation. As the compound that is to be mapped onto a fragment hit may contain differences, primarily as expansions, relative to the reference compound, multiple round of maximum common substructure mapping are employed. First a strict mapping is performed, then a series of more lax mappings are performed constrained on the first mapping in order to cover the most atoms, while still maintaining the core details. When mapping a compound to multiple hits, the mappings covering the largest number of atoms is likewise used to constrain the mapping of the other hits.

2.2 Combination benchmark

Combinations on test datasets were conducted to assess success rate and availability from make-on-demand space. The hits from the XChem targets SARS-COV-2 MPro (cysteine protease)[17] and Mac1 domain of SARS-COV-2 NSP3 (macrodomain ADP-ribosylhydrolase)[18], were downloaded from Fragalysis (https://fragalysis.diamond.ac.uk/) and filtered for inclusion in the DSi-Poised library [19]. The templates used were PDB:6LU7 for MPro and PDB:6WOJ for Mac1, these were energy minimised with PyRosetta with the FastRelax mover constrained by its density-map[15]. Their hits were combined (merged/linked) in order to quantify the failure rate and the synthetic accessibility. Additionally, to explore the thermodynamic cost of fidelity to the reference compounds, alternative approaches were adopted, namely merging solely by maximum common substructure and merging by BRICS decomposition[20]. These were placed with the PyRosetta framework of Fragmenstein (Igor). BREED[21] with 1.5 Å cut-off and with the "untangle" setting disabled was also run, but the limited results precluded its benchmarking. Interactions were determined with PLIP [22]. Interactive pages of results were created in MichelaNGLo [23].

2.3 Placement benchmarking

MPro was used to assess the accuracy of placements of follow-up compounds. The information of which fragment hits were inspirations for which crystallised follow-up compounds was taken from the Moonshot GitHub repository[24], but was reduced to contain only the relevant inspiration hits for each submitted compound as these are presented together for each submission set. Namely, the relevant hits were manually picked based on the binding of the hits and the 2D representation of the follow-up in order to not bias the selection (*cf,* code in repository). The common protein template used was PDB:6LU7, which was minimised as describe above. [15]Fragmenstein was run with the correction that the PyRosetta pose instance was modified to have catalytic His41 protonated on Nδ (HID) and Cys145 deprotonated for non-covalent compounds, while for compounds with electrophilic warheads His41 protonated on Nε (HIE) and Cys145 crosslinked with the compound. The latter functionality is automatic if the SMILES to be placed has a dummy atom (* in SMILES) or the warhead conversion code within Fragmenstein is called.

RDock[25] was executed on the same Mpro merges that were placed with Fragmenstein. For each compound, the protein cavity was defined using the RbtLigandSiteMapper on the largest inspirational fragment hit with a radius of 8Å and the following parameters: SMALL_SPHERE 1.0; MIN_VOLUME 100; MAX_CAVITIES 1; VOL_INCR 0.0; and GRIDSTEP 0.5.

One hundred poses per compounds were docked using the default “dock.prm” protocol. The top poses were selected based on the rDock score and the best RMSDs.

For the case of constrained docking, we computed the pharmacophores of the hits and set them as optional restraints with weight 1. The percent of constraints that should be satisfied was set to 80% based on a preliminary calibration test to achieve the lowest RMSD from the crystallographic pose. It is importance to notice that in a real-world scenario this calibration strategy is not possible since the crystallographic poses are not available, consequently, the results here presented are an overestimation of the actual performance.

2.4 Case examples

Two special examples were retrospectively analysed, specifically addressing covalent ligands and the user-provided mapping.

To demonstrate the need for the thermodynamic corrections in the final step of Fragmenstein, the placement of two follow-up compounds binding NUDT7 from [26] (deposition group G_1002045) were investigated. NU181 (PDB:5QH1, chemical component: H5G, Enamine: Z1632454068) and PCM-0102716 (PDB: 5QH9, chemical component: GZY, Enamine: Z254513422) were the inspiration hits for NU443 (PDB: 5QHH, chemical component: H5D, S enantiomer) and NU442 (PDB:5QHG, chemical component: H17, R enantiomer), which were modelled with the chloroacetamide reacted with Cys73.

To demonstrate the use of user correction, the placement of the follow-up compound binding the tubulin interface from [27] (deposition groups G_1002173 and G_1002214) was investigated. F04 (PDB: 5S4O, chemical component: O0J, Enamine: Z48847594) and F36 (PDB: 5S5K, chemical component: S6V, Enamine: Z2472938267) were the inspiration hits for todalam-4 (PDB: 5SB3, chemical component: 47F, Enamine: Z48853939). The modelling was done with a custom map in order to flip the N and S atoms in the aminothiazole, a equally plausible orientation given the electron density and required for the elaboration.

3. Results

3.1 Combination benchmark

On two datasets, Fragmenstein proposes 49 and 24 easily accessible follow-up compounds from the combination of 34 and 44 inspiration hits. Fragmenstein merges fragments by first stitching together the positioned atoms of the inspiration fragments prior and then locally minimising under strong constraints, without relying on previously generated conformers. To assess the overall quality of combinations from Fragmenstein, *i.e.* determining the methodological failure rate and synthetic accessibility, two targets, MPro (a cysteine protease from SARS-COV-2), Mac1 (a nucleosyl-peptide hydrolase from SARS-COV-2) from previous fragment screens were chosen and the initial hits that originating from a library designed to facilitate synthetic follow-ups (DSi-Poised) were combined and scored (Table 1, interactive at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-mpro-DSiP). Excluding the combinations that were over 5Å apart for their closest atoms, the failure rate was 1.4% due to compounds whose chemistry could not be rectified correctly, while 56% of combinations were energetically favourable ($\Delta\Delta G < 0$) without excessive deviation from the positions of the inspiration hits (RMSD < 1.). Of the 420 acceptable combinations, 7 were purchasable, while 64 could potentially be made with 2 or fewer reactions according to predictions from PostEra Manifold. Therefore, Fragmenstein was able to suggest a constructive number of synthetically accessible compounds that are predicted to strongly follow the binding conformation of the inspiration fragment hits, which is the aim of Fragmenstein.

The strict obedience to atomic positions by Fragmenstein is a strong filter whose effects may be mislead by potentials and are unmasked when counting number of interactions. In fact, a key point of Fragmenstein is obedience to inspiration hits, which is not shared by other methods, therefore, warranting further investigation. To emphasise the importance of fidelity to position of the initial hits of the largest and worst performing set (Mac1) were combined ignoring positional information, by either merging by maximum common substructure (MCS) and by BRICS decomposition, and additionally merged with Fragmenstein but constrained to a single hit.

The minimisation of these in place via constraints to both the inspiration hits did not yield any acceptable poses, whereas misleadingly the minimisation in place against only the larger hit resulted in a jump to 23% for MCS and 34% for BRICS (Figure S3). When Fragmenstein mergers were constrained to a single hit, the acceptance rate increased from 11% to 14%, because several mergers that were irreconcilably strained when constrained against two hits were more relaxed when constrained against a single hit and not obliged to respect the position of the second hit.

The number of interactions formed as determined via PLIP reveals a median 0.25 interactions per heavy-atom (HA) for the acceptable two-hit–constrained Fragmenstein mergers and a lower 0.21 interactions/HA for single-hit–constrained Fragmenstein mergers. Thus unmasking the cost of having fewer constraints.

This is because without the positional constraints the physics of

the force-field start to dominate the placement, which may be imperfect. Fragmenstein utilises molecular mechanics, but does not find the energy minimum within a box, and instead finds a low energy state around the initial hit. As a consequence, the calculated free energy of binding are sensitive to the number of constraints applied and are not an overly meaningful metric. Unsurprisingly the median ligand efficiency misleadingly improves from −0.20 kcal/mol/HA for the two-hit–constrained mergers to −0.23 kcal/mol/HA for the single-hit–constrained mergers, despite the latter forming less meaningful interactions by not obeying the conformation of the second hit.

The pure-MCS mergers constrained to the largest hit had both fewer interactions and worse free energy of binding (median ligand efficiency of –0.14 kcal/HA) due to the more compact nature, making the mergers more likely to fall off an energy cliff. This is in contrast to BRICS decomposition where the substructures of the inspiration hits are joined at the broken bonds therefore respecting the axis of the compounds, even if they may not have been spatially overlapping. In the BRICS approach, the constraints were to a substructure of single hit, so the ligand efficiency is misleadingly better than Fragmenstein (−0.25 kcal/mol/HA), whereas the median number of interactions was actually lower (0.17 interactions/HA). This is a reminder that the equations and parameters in force-fields used for docking are not fully accurate [4], and the sole reliance on these can be highly misleading.

	MPro	Mac1
Number of hits used	34	44
Number of acceptable[a] mergers	157	263
Number of failed mergers due to equal size to one hit	13	34
Number of failed mergers due to > 5 Å minimum distance between hits	918	1438
Number of failed mergers due to strain (ΔΔG > 0 kcal/mol or >1 Å RMSD)	33	149
Number of failed mergers due to technical issues	1	8
median mol. wt of acceptable subset	356.1	305.0
median QED [b] of acceptable subset	.79	.66
Number of of acceptable compounds with SA [c] < 0.	54	27
Number of of acceptable compounds with SA≤0.4	71	40
Number of acceptable compounds that are purchasable[d]	5	2
Number of acceptable compounds with purchasable analogues in Enamine Real differing by 2 or fewer atoms	26	22
Number of acceptable compounds accessible via a one-step synthesis[e]	28	10
Number of acceptable compounds accessible via a two-step synthesis	16	10

Table 1. Assessment of the quality of mergers generated with Fragmenstein. Combinations (mergers/Linkages) were computed for DSiPoised subset of hits for the targets and classified by outcome and then the acceptable molecules were further assessed for synthentic accessibility.

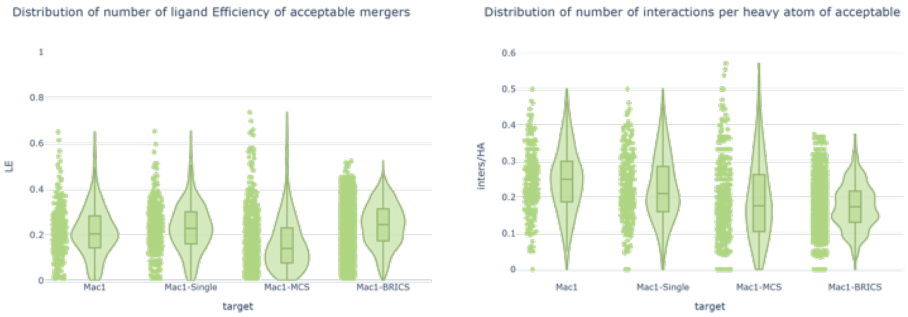
a) The acceptability criteria were both hits were used, RMSD < 1 Å, ΔΔG > 0 kcal/mol, and number of heavy atoms greater than that of the largest. hit,

b) QED: Quantitative Estimate of Druglikeness, calculated by RDKit.

C) SA: FastSAScore calculated by Postera Manifold.

D) Purchasable: compound available from the vendors Enamine (BB, MADE and REAL), Sigma, Mcule, EMolecules, Molport, WuXi (BB and GalaXi).

E) 1-step / 2-step: Molecule unavailable but synthesisable in a one or two reactions as predicted by by Postera Manifold retrosynthesis. The combinations can be inspected at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-mpro-DSiP.



Supplementary figure 3. Distribution of ligand efficiency (left) and of number of interactions per heavy atom for the different merger performed on the Mac1 poised dataset, namely Fragmenstein, Fragmenstein modified to be constrained to a single hit, MCS merger (void of positional information) and BRICS decomposition and building. The median (centre line within each the box) for BRICS and for the single-hit–constrained Fragmenstein are larger than the two-hit–constrained Fragmenstein median in the ligand efficiency plot, but not in the interaction count plot because the minimisation for the regular Fragmenstein is more constrained thus affecting the calculated Gibbs free energy of binding.

3.2 Placement benchmark

A retrospective placement of 100 compounds based on their inspiration reveals much strong agreement with the crystal structures than pharmacophoric constraints. The governing idea behind Fragmenstein is striving for fidelity to the position of the inspiring hits based upon the assumption that the follow-up compounds bind in a very similar way. To test this hypothesis the dataset from the Covid Moonshot project was used as a large panel of hit-inspired follow-up compounds is available [24]. This was a collaborative SAR-COV-2 protease fragment-based drug discovery project that relied on user submitted follow-up ideas. These submissions were guided by user's choice and as a result represent a spectrum of diverse approaches. The submissions were filtered for compounds that were crystalised and that had a stated inspiration, yielding a total of 100 compounds. The atomic positions of the conformer from the crystal structure were compared to those of a conformer placed by Fragmenstein constrained against the stated inspiring hits and to those of conformers docked with and without restraints (Figure 2, interactive at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-moonshot).

The importance of exploiting the structural information of the inspirational hits is illustrated by the fact that 64% of the proposed merges were found to fully preserve the pose of their inspirational fragments (mean RMSD<2Å).

Fragmenstein was able to propose high-quality poses (RMSD < 1Å) for 28% of the evaluated compounds and acceptable poses (RMSD < 2Å) for 56% of them. On the other hand, docking (via rDock) was not able to obtain any high-quality poses (Figure 2A).

In order to determine if Fragmenstein was able to better exploit the structural information of the fragment hits than other approaches, we next compared Fragmenstein with the constrained version of rDock using pharmacophoric constraints derived from the inspirational hits. Figure 2B shows that, while including constraints improves the docking performance, Fragmenstein still outperforms rDock, which is able to produce accurate poses for only 5% of the compounds. A factor involved is that Fragmenstein generates the conformer based on the hits, while docking frequently choses a conformer among a set of generated conformers. Specifically, for this dataset, the most similar generated conformers out of 10, 100 and 1,000 (ETKDG in RDKit) to the crystallographic pose deviated by 0.9 Å, 0.7Å and 0.6 Å on average. This encapsulating the reason underlying the choice in Fragmenstein to start from a stitched-together conformer. This together with the hit-derived strong constraints during minimisation allows the placed molecule to be highly faithful to the inspiration hits.

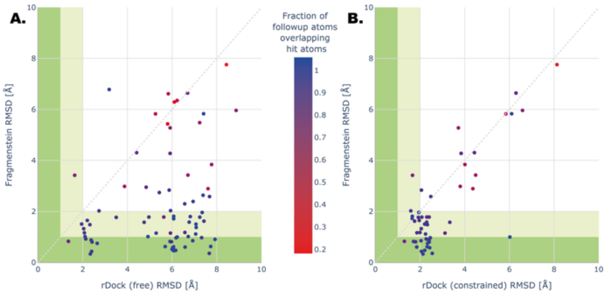


Figure 2. Accuracy of placement of Covid19 MPro1 Moonshot compounds. Follow-up compounds in the Covid19 MPro1 Moonshot project which had a stated inspiration (manually adjusted) were placed with Fragmenstein and docked with rDock either freely or with pharmacophore constraints. The initial dataset contained 100 fragment-inspired compounds, but 23 were discarded (because the crystal structure of follow-up had no overlapping atoms with the inspiration, the reactive follow-up was non-covalent in the crystal structure and/or Fragmenstein failed to minimise the follow-up compound) and a further 20 were discarded in the pharmacophore constrained rDock due to failure to dock successfully. Green area < 1 Å RMSD against crystal structure, pale green < 2 Å RMSD. The compounds bound in the same pocket as the hits but the Fragmenstein models had an RMSD > 5 Å were x2581, x10236, x2764, x10900, x2779, x1386, x3305, x1384, x10606, x10723, x10049, x3366, for most of these either the crystallised compound disobeyed the hits or Fragmenstein incorrectly mapped the follow-up to the hits due the convoluted overlay. Individual models can be investigated at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-moonshot.

3.3 Case examples

Fragmenstein can not only work with non-covalent compounds, but also with covalent compounds. In order to work with covalent compounds, Fragmenstein treats the attachment atom (stored as a dummy atom) and defined atoms from the warhead differently, primarily by protecting these during merging. To test the impact of having a covalent attachment, the placement of a published compound [26] with two stereoisomers was replicated. In this study, only one enantiomer reacted with the thiol of the catalytic cysteine in the protein (NUDT7).

This compound is merger of two hits (NU181 and PCM0102716) which were used for placement with Fragmenstein. The RMSD between the placed model and the crystal structure of the merger is 0.28 Å, while the combined RMSD values of the model and the structure against the inspiration hits are 0.65 and 0.61 Å, indicating that the slight conformational change resulting from the constrained minimisation is also seen in the crystal structure. This placement (Figure 3A) operation also showcases a feature of Fragmenstein borne out of having to operate on multiple hits. Namely, that some superposed substituents in the hits may act as red herrings and are ignored, in this example the hydroxyl of one hit (NU181) is automatically ignored from the mapping as it would otherwise impede the mapping of the second hit (PCM0102716) which has a group occupying the same space. In this fragment screens, as is common, a racemic mix first soaked in the crystal (NU308) and was subsequently chirally separated into two stereoisomers (NU442 and NU443). Whereas one stereoisomer (NU443) was found covalently bound, the other (NU442) was found not reacted. Placing with Fragmenstein the latter stereoisomer as a covalent compound (Supplementary figure 4), yielded a pose with a 10% worse binding ΔG (predicted by Rosetta ref2015 scorefunction without constraint weights) than the former and with a 0.9 Å shift in the sulfur atom of the connected cysteine relative to the position in the inspiration hit, indicating that the covalent bond is expected to be strained as is confirmed in the crystal structure wherein the presumably worse reaction barrier was not overcome.

In Fragmenstein, it is possible to enforce certain follow-up atoms to map to specific atoms from the hit atoms in order to get the intended placement. An example of this is an inspiration hit with a ring in a flipped conformation. Crystallographic structures generally consist of a single conformer bound in a set orientation as suggested by the electron density map. In some cases, for example with the terminal amides of glutamine or asparagine or the ring in a histidine the specific density alone cannot reveal which way these sidechains are oriented. This may also apply to ligands.

An example of this is seen with tubulin inhibitor Todalam-4 [28]. This compound is the merger of two compounds (F04 and F36). One possesses an aminothiazole ring placed in one orientation in the crystal structure, yet for the merger to be accurate, the flipped orientation is required. When applied to this test case, when passed a map to override certain atoms Fragmenstein correctly predicts the intended placement (Figure 3B). This ability to fine tune the behaviour of Fragmenstein allows it to be highly versatile and adaptable.

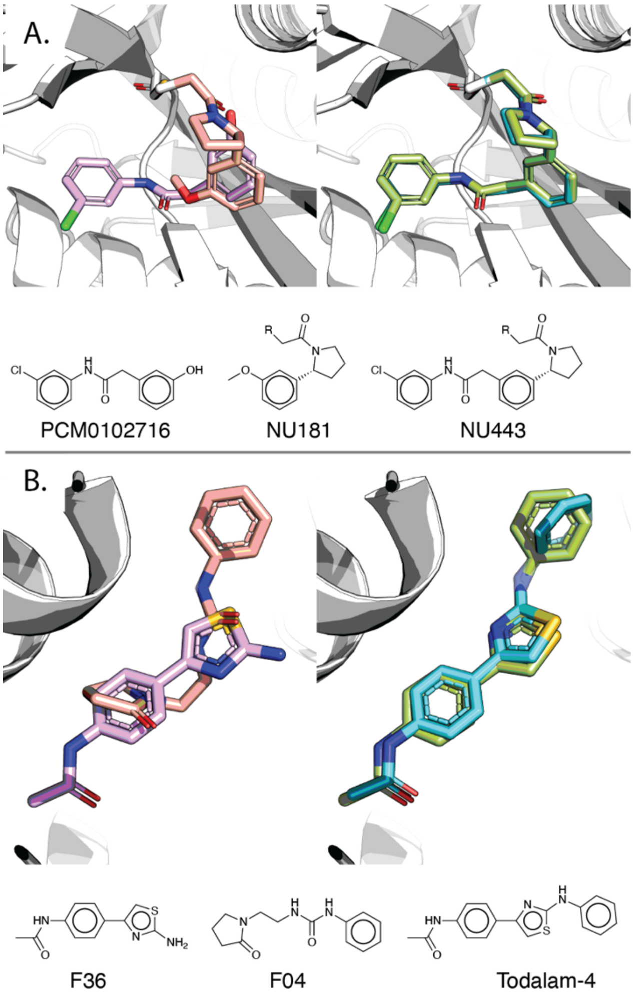
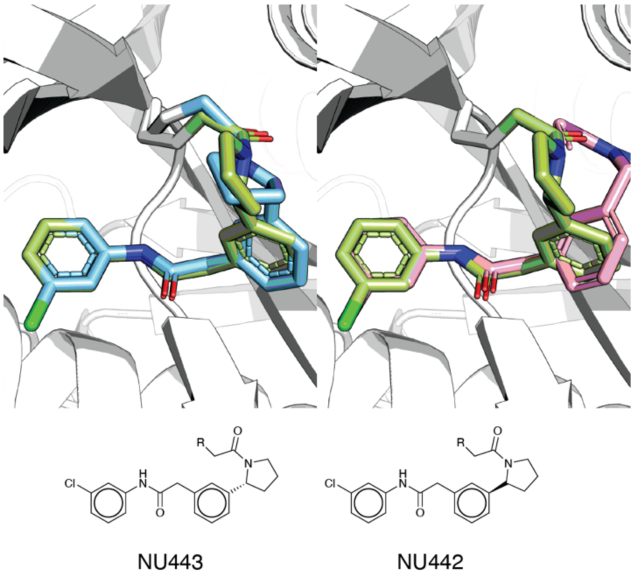


Figure 3. Retrospective Comparison of crystallised and placed follow-up compound from NUDT7 study (A) and tubulin (B) study, illustrating an merger with hits that do not overlap cleanly and a merger requiring a user-defined mapping respectively.

In the NUDT7 study, the two hits NU181 (in lavender, LHS) and PCM0102716 (puce, LHS) were merged in <paper> yielding the merger NU443. The crystal structure of NU443 (turquoise, RHS) overlayed with the placement predicted by Fragmenstein (green, RHS). PCM0102716 and NU443 are covalent with Cys73 via a acryloyl warhead. Internally outside of the PyRosetta operations, covalent attachment atoms are stored as dummy/R/✱ atoms, shown in white. The RMSD between the placed model and the crystal structure of NU443 is 0.28 Å, while the combined RMSD values of the model and the structure against the inspiration hits are 0.65 and 0.61 Å. In the placement process the hydroxyl of NU181 was automatically discarded from the mapping as it would otherwise impede the mapping of the second hit (PCM0102716) due to the greater proximity of the NU181 hydroxyl to the oxygen of the acryloyl warhead of PCM0102716 rather than to the carbon bonded the benzene ring in PCM0102716.

In the tubulin study, F04 (purple) and F36 (orange) inspired Todalam-4 (skyblue: crystal, green: predicted). The aminothiazole ring is flipped between F04 and todalam-4 by design. A constructive observation of this follow-up is that the N-benzyl is rotated in the crystal relative to F36 possibly to attain a T-shaped pi bond, a dipole-momentum–driven configuration, which is not modelled in classical mechanics forcefields such as that employed by Rosetta.



Supplementary Figure 4. Placement of NU442. Turquoise: hits, green: Fragmenstein-generated conformer (covalent), puce: crystallographic conformer (unreacted). The sulfur atom is shifted in the Fragmenstein conformer relative to the inspiration due to the strain imposed by the chirality of the pyrollidine substituent.

4. Discussion

4.1 Rationale

Strict adherence to inspiration hits in follow-up design aids human assessment. The core principle of Fragmenstein is to create a conformer of a compound, via its two routes (combinations or placements) by stitching together the atomic positions of the inspiration hits, with the aim of being as faithful as possible to these without being energetically unfeasible. A prime reason is to maximise the confidence in the solution by the appraising human irrespective of metrics.

Thanks to recent computational advances, there has been a move towards increased automation in both compound design and selection, but nevertheless a synergistic approach between human and machine has been described as a key factor in a successful drug discovery campaign [29]. In the MPro Covid19 Moonshot project, this was seen, where the lead was derived from a manually designed submission (TRY-UNI-714a760b-6) and textual analysis of the descriptions in the user submissions from the Moonshot project revealed that hypothesis-driven compounds had better pIC50 values [24]. This emphasises the need for human appraisal. Human appraisal becomes more and more critical in latter steps of a drug discovery campaign as one is generally severely curtailed by chemical space and as a result these steps become heavily human driven. A key part of human assessment is visual inspection. When the follow-up compound is docked in a different way from the inspiring hits it may become hard to assess its contacts, especially as this may not be what may be seen in the eventual crystal structures, therefore having a predicted conformation positioned faithfully to the inspirations, which Fragmenstein can provide, becomes highly advantageous.

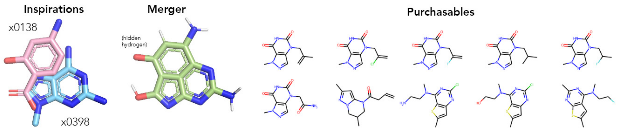
4.2 Merging

Physics-based scores can mislead the user if the atomic positions of inspiration hits are not obeyed. As seen in table 1 Fragmenstein has a very high success rate and yields several compounds in made-on-demand space. Combining compounds, or their substructures with high group efficiency, yields novel compounds. Fragmenstein aims to preserve the interactions of the inspiration hits unlike other methods. Comparing these, one may be misled by the metrics used. Gibbs free energy of binding can be misleading, especially when constraints are involved: reducing the number of constraints improves this metric, whereas the number of interactions is lower. As discussed, redocking tests do not have a high success rate. This is because static snapshot scores from simplified molecular thermodynamics can be imperfect due to a plethora of reasons, such as incorrect π–π orientations due to absent dipole moments as seen in Figure S5. Hence the need for positional fidelity and caution when assessing these metrics.

Fragmenstein can work with overlapping compounds. The linking approach is intentionally basic as Fragmenstein is not intended for Protac design (*i.e.* two distinct moieties tethered by a long flexible linker) or to add novel chemical substructures between two hits. These use cases are addressed by other tools. However, for close compounds, the torsion of the link can be expected to be highly constrained by the inspiration hits, which is exactly the sort of problem Fragmenstein is meant to address as demonstrated in its role in the identification of a IC_{50} 430 nM inhibitor against SARS-COV-2 Mac1[18,30] (mergers: https://michelanglo.sgc.ox.ac.uk/r/fragmenstein_nsp3).

Even though the compounds generated by combination are chemical correct, a limitation of this is that the compounds created may not be in make-on-demand space or may not be synthetically accessible. In the provided demonstration notebook the SmallWorld server is queried to find purchasable analogues from Enamine REAL, which can be placed by Fragmenstein. A similar approach was used in the SARS-COV-2 Mac1 study [18] (using Arthor). Chemical make-on-demand space despite its vastness is often limiting. In fact, it should be noted that the outcome of the search may not be always fruitful. For example, a merger of two perfectly placed inspirations may yield a compound that is far removed from make-on-demand space (e.g. Supplementary figure 5, a clear merger distant from make-on-demand space), thus forcing the user to consider other mergers or linkers as a starting point for exploration. Predictably the more the lead-like candidates grow, the more isolated they may be in easily synthesisable chemical space and more skirting around is required to address issues present.

A fruitful synergism to optimise compounds is combing BRICS decomposition and Fragmenstein, which in effect removes substructures from the initial hits which are not forming good interactions or hamper synthetic accessibility.



Supplementary Figure 5. Example of legitimate merger from Mac1, wherein the acenaphthylene core is chemically sound, but for which no analogues are present in make-on-demand space.

4.3 Placement

Follow-up compounds do not always obey the inspiration hits, but when they do they frequently follow them strictly to the benefit of the user. Docking is often employed to rank a large number of compounds, which were chosen via one of many possible approaches leveraging the fragment hits to differing extents. However, docking has the problem that the outputted conformer may not reflect the binding of the fragment hits that inspired them, even though fragment hits with a common substructure are most often found positioned in a very similar manner (*vide supra*). Were a docked follow-up candidate to interact differently than its inspirations, the validity of its score would be rightfully put to question by an experimentalist. Pharmacophoric constraints and hotspot mapping partially addresses this, but still falls short by allowing deviation from the core of the inspiration hits. Several decomposition studies address the SAR additivity/superadditivity of certain functional groups[31–34] and how the conformation is maintained crystallographically. Herein, the opposite direction is taken and is found to be also consistent. Fragmenstein, aims to help assess the credibility of a given compound. In Figure 2 it was shown that in the Covid-Moonshot dataset of the crystallised follow-up compounds that bound similarly to their inspiration (69%), 82% are placed with an RMSD under 2Å compared to 22% by pharmacophore-constrained docking. Confirming the importance of obeying the position of the atoms in the inspiration hits.

Fragmenstein can be combined with compound searches for effective enumeration of candidate follow-ups. In addition to the three-step route of combining hits, searching for their purchasable analogues in an ultra-large library and placing these, a recent published approach for fragment joining enumerate all purchasable compounds that contain substructure of pairs of hits and places these with Fragmenstein[35]. This joining approach addresses the limitation of the combination step of Fragmenstein when dealing with hits that are separated by several ångströms and addresses the problem of conformationally filtering the shortlisted compounds thanks to Fragmenstein by enforcing that these obey the hit substructures.

4.4 Multistep aid

Fragmenstein can help at multiple and at different steps along the way. Beyond drug discovery, Fragmenstein has found some uses in biochemistry settings by virtue of allowing the change of an crystallographically amenable analogue with the native substrate, *e.g.* the non-hydrolysable guanosine imidotriphosphate (GNP) for guanosine triphosphate (GTP) [36]. A novel unexplored use of the placement feature in Fragmenstein could be to explore the native substrate, transition state or product of an enzyme. These binding modes often get rediscovered in drug discovery campaigns . Therefore, drug discovery campaigns could leverage the knowledge of native substrates and intermediates in order to gain specificity or robustness against resistance.

5. Conclusions

Fragmenstein is first and foremost tool that strictly obeys the inspiration hits both as a generative model and as a docking alternative. This provides a way for a human user to drive their computational experiment to meet their hypothesis by controlling and appraising the prediction: in the end, the decision of which compounds to purchase is very rarely left to a blind algorithm and instead is put in the hands of an experienced but often diffident chemist. Fragmenstein by strictly obeying the empirical data can provide the sought reassurance.

Author Contributions

MP. Ferla: Conceptualization, Methodology, Software, Data Curation, Validation, Writing - Original Draft. **R. Sánchez-García**: Validation. **RE. Skyner**: Conceptualization, Methodology, Writing - Original Draft. **S. Gahbauer**:. **BD. Marsden**: Supervision. **JC. Taylor**: Supervision, Funding acquisition. **CM. Deane**: Supervision, Conceptualization, Validation, Writing - Review & Editing, Funding acquisition. **F. von Delft**: Conceptualization, Funding acquisition

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work was in part supported by the National (UK) Institute for Health Research (NIHR) Oxford Biomedical Research Centre (BRC), the Wellcome Trust Core Award [203141/Z/16/Z], Rosetrees Trust award [M940], and National (US) Institutes of Health (NIH) through the NIAID Antiviral Drug Discovery (AViDD) U19 Program [U19AI171399].

References

1 D. A. Erlanson, *Top Curr Chem*, 2012, **317**, 1–32.

2 L. R. de Souza Neto, J. T. Moreira-Filho, B. J. Neves, R. L. B. R. Maidana, A. C. R. Guimarães, N. Furnham, C. H. Andrade and F. P. Silva, *Front Chem*, 2020, 8, 93.

3 P. H. M. Torres, A. C. R. Sodero, P. Jofily and F. P. Silva-Jr, *Int J Mol Sci*, 2019, **20**, 4574.

4 D. R. Houston and M. D. Walkinshaw, *J Chem Inf Model*, 2013, **53**, 384–390.

5 P. R. Curran, C. J. Radoux, M. D. Smilova, R. A. Sykes, A. P. Higueruelo, A. R. Bradley, B. D. Marsden, D. R. Spring, T. L. Blundell, A. R. Leach, W. R. Pitt and J. C. Cole, *J Chem Inf Model*, 2020, **60**, 1911–1916.

6 T. Liu, M. Naderi, C. Alvin, S. Mukhopadhyay and M. Brylinski, *J Chem Inf Model*, 2017, **57**, 627–631.

7 J. O. Spiegel and J. D. Durrant, *Journal of Cheminformatics 2020 12:1*, 2020, **12**, 1–16.

8 F. Imrie, A. R. Bradley, M. Van Der Schaar and C. M. Deane, *J Chem Inf Model*, 2020, **60**, 1983–1995.

9 F. Dey and A. Caflisch, *J Chem Inf Model*, 2008, **48**, 679–690.

10 F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *Chem Sci*, 2021, **12**, 14577–14589.

11 T. E. Hadfield, F. Imrie, A. Merritt, K. Birchall and C. M. Deane, *J Chem Inf Model*, , DOI:10.1021/ACS.JCIM.1C01311.

12 A. C. Pierce, G. Rao and G. W. Bemis, *J Med Chem*, 2004, **47**, 2768–2775.

13 P. O. Nikiforov, S. Surade, M. Blaszczyk, V. Delorme, P. Brodin, A. R. Baulard, T. L. Blundell and C. Abell, *Org Biomol Chem*, 2016, **14**, 2318–2326.

14 S. Riniker and G. A. Landrum, *J Chem Inf Model*, 2015, **55**, 2562–2574.

15 S. Chaudhury, S. Lyskov and J. J. Gray, *Bioinformatics*, 2010, **26**, 689.

16 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J Chem Inf Model*, 2020, **60**, 6065–6073.

17 T. C. M. Consortium, H. Achdout, A. Aimon, D. S. Alonzi, R. Arbon, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, V. A. Bilenko, V. A. Bilenko, M. L. Boby, B. Borden, P. Boulet, G. R. Bowman, J. Brun, L. Brwewitz, S. Bvnbs, M. Calmiano, A. Carbery, D. Carney, E. Cattermole, E. Chang, E. Chernyshenko, J. D. Chodera, A. Clyde, J. E. Coffland, G. Cohen, J. Cole, A. Contini, L. Cox, T. I. Croll, M. Cvitkovic, A. Dias, K. Donckers, D. L. Dotson, A. Douangamath, S. Duberstein, T. Dudgeon, L. Dunnett, P. K. Eastman, N. Erez, C. J. Eyermann, M. Fairhead, G. Fate, D. Fearon, O. Fedorov, M. Ferla, R. S. Fernandes, L. Ferrins, M. Filep, R. Foster, H. Foster, L. Fraisse, R. Gabizon, A. Garcia-Sastre, V. O. Gawriljuk, P. Gehrtz, C. Gileadi, C. Giroud, W. G. Glass, R. Glen, I. Glinert, A. S. Godoy, M. Gorichko, T. Gorrie-Stone, E. J. Griffen, S. Hahn, A. Haneef, S. H. Hart, J. Heer, M. Henry, M. Hill, S. Horrell, Q. Y. Huang, V. D. Huliak, V. D. Huliak, M. F. D. Hurley, T. Israely, A. Jajack, J. Jansen, E. Jnoff, D. Jochmans, T. John, S. De Jonghe, B. Kaminow, L. Kang, A. L. Kantsadi, P. W. Kenny, J. L. Kiappes, S. O. Kinakh, S. O. Kinakh, L. Koekemoer, B. Kovar, T. Krojer, V. La, A. A. Lee, B. A. Lefker, H. Levy, I. G. Logvinenko, I. G. Logvinenko, N. London, P. Lukacik, H. B. Macdonald, E. M. MacLean, L. L. Makower, T. R. Malla, T. Matviiuk, W. McCorkindale, B. L. McGovern, S. Melamed, K. P. Melnykov, K. P. Melnykov, O. Michurin, P. Miesen, H. Mikolajek, B. F. Milne, D. Minh, A. Morris, G. M. Morris, M. J. Morwitzer, D. Moustakas, C. Mowbray, A. M. Nakamura, J. B. Neto, J. Neyts, L. Nguyen, G. D. Noske, V. Oleinikovas, G. Oliva, G. J. Overheul, D. Owen, R. Pai, J. Pan, N. Paran, A. Payne, B. Perry, M. Pingle, J. Pinjari, B. Politi, A. Powell, V. Psenak, I. Pulido, R. Puni, V. L. Rangel, R. N. Reddi, P. Rees, S. P. Reid, L. Reid, E. Resnick, E. G. Ripka, M. C. Robinson, R. P. Robinson, J. Rodriguez-Guerra, R. Rosales, D. A. Rufa, K. Saar, K. S. Saikatendu, E. Salah, D. Schaller, J. Scheen, C. A. Schiffer, C. Schofield, M. Shafeev, A. Shaikh, A. M. Shaqra, J. Shi, K. Shurrush, S. Singh, A. Sittner, P. Sjo, R. Skyner, A. Smalley, B. Smeets, M. D. Smilova, L. J. Solmesky, J. Spencer, C. Strain-Damerell, V. Swamy, H. Tamir, J. C. Taylor, R. E. Tennant, W. Thompson, A. Thompson, S. Tomasio, C. Tomlinson, I. S. Tsurupa, I. S. Tsurupa, A. Tumber, I. Vakonakis, R. P. van Rij, L. Vangeel, F. S. Varghese, M. Vaschetto, E. B. Vitner, V. Voelz, A. Volkamer, F. von Delft, A. von Delft, M. Walsh, W. Ward, C. Weatherall, S. Weiss, K. M. White, C. F. Wild, K. D. Witt, M. Wittmann, N. Wright, Y. Yahalom-Ronen, N. K. Yilmaz, D. Zaidmann, I. Zhang, H. Zidane, N. Zitzmann and S. N. Zvornicanin, *bioRxiv*, 2023, 2020.10.29.339317.

18 S. Gahbauer, G. J. Correy, M. Schuller, M. P. Ferla, Y. U. Doruk, M. Rachman, T. Wu, M. Diolaiti, S. Wang, R. J. Neitz, D. Fearon, D. S. Radchenko, Y. S. Moroz, J. J. Irwin, A. R. Renslo, J. C. Taylor, J. E. Gestwicki, F. von Delft, A. Ashworth, I. Ahel, B. K. Shoichet and J. S. Fraser, *Proc Natl Acad Sci U S A*, 2023, **120**, e2212931120.

19 O. B. Cox, T. Krojer, P. Collins, O. Monteiro, R. Talon, A. Bradley, O. Fedorov, J. Amin, B. D. Marsden, J. Spencer, F. Von Delft and P. E. Brennan, *Chem Sci*, 2016, **7**, 2322–2330.

20 J. Degen, C. Wegscheid-Gerlach, A. Zaliani and M. Rarey, *ChemMedChem*, 2008, **3**, 1503–1507.

21 A. C. Pierce, G. Rao and G. W. Bemis, *J Med Chem*, 2004, **47**, 2768–2775.

22 S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme and M. Schroeder, *Nucleic Acids Res*, 2015, **43**, W443–W447.

23 M. P. Ferla, A. T. Pagnamenta, D. Damerell, J. C. Taylor and B. D. Marsden, *Bioinformatics*, 2020, **36**, 3268–3270.

24 T. C. M. Consortium, H. Achdout, A. Aimon, D. S. Alonzi, R. Arbon, E. Bar-David, H. Barr, A. Ben-Shmuel, J. Bennett, V. A. Bilenko, V. A. Bilenko, M. L. Boby, B. Borden, P. Boulet, G. R. Bowman, J. Brun, L. Brwewitz, S. Bvnbs, M. Calmiano, A. Carbery, D. Carney, E. Cattermole, E. Chang, E. Chernyshenko, J. D. Chodera, A. Clyde, J. E. Coffland, G. Cohen, J. Cole, A. Contini, L. Cox, T. I. Croll, M. Cvitkovic, A. Dias, K. Donckers, D. L. Dotson, A. Douangamath, S. Duberstein, T. Dudgeon, L. Dunnett, P. K. Eastman, N. Erez, C. J. Eyermann, M. Fairhead, G. Fate, D. Fearon, O. Fedorov, M. Ferla, R. S. Fernandes, L. Ferrins, M. Filep, R. Foster, H. Foster, L. Fraisse, R. Gabizon, A. Garcia-Sastre, V. O. Gawriljuk, P. Gehrtz, C. Gileadi, C. Giroud, W. G. Glass, R. Glen, I. Glinert, A. S. Godoy, M. Gorichko, T. Gorrie-Stone, E. J. Griffen, S. Hahn, A. Haneef, S. H. Hart, J. Heer, M. Henry, M. Hill, S. Horrell, Q. Y. Huang, V. D. Huliak, V. D. Huliak, M. F. D. Hurley, T. Israely, A. Jajack, J. Jansen, E. Jnoff, D. Jochmans, T. John, S. De Jonghe, B. Kaminow, L. Kang, A. L. Kantsadi, P. W. Kenny, J. L. Kiappes, S. O. Kinakh, S. O. Kinakh, L. Koekemoer, B. Kovar, T. Krojer, V. La, A. A. Lee, B. A. Lefker, H. Levy, I. G. Logvinenko, I. G. Logvinenko, N. London, P. Lukacik, H. B. Macdonald, E. M. MacLean, L. L. Makower, T. R. Malla, T. Matviiuk, W. McCorkindale, B. L. McGovern, S. Melamed, K. P. Melnykov, K. P. Melnykov, O. Michurin, P. Miesen, H. Mikolajek, B. F. Milne, D. Minh, A. Morris, G. M. Morris, M. J. Morwitzer, D. Moustakas, C. Mowbray, A. M. Nakamura, J. B. Neto, J. Neyts, L. Nguyen, G. D. Noske, V. Oleinikovas, G. Oliva, G. J. Overheul, D. Owen, R. Pai, J. Pan, N. Paran, A. Payne, B. Perry, M. Pingle, J. Pinjari, B. Politi, A. Powell, V. Psenak, I. Pulido, R. Puni, V. L. Rangel, R. N. Reddi, P. Rees, S. P. Reid, L. Reid, E. Resnick, E. G. Ripka, M. C. Robinson, R. P. Robinson, J. Rodriguez-Guerra, R. Rosales, D. A. Rufa, K. Saar, K. S. Saikatendu, E. Salah, D. Schaller, J. Scheen, C. A. Schiffer, C. Schofield, M. Shafeev, A. Shaikh, A. M. Shaqra, J. Shi, K. Shurrush, S. Singh, A. Sittner, P. Sjo, R. Skyner, A. Smalley, B. Smeets, M. D. Smilova, L. J. Solmesky, J. Spencer, C. Strain-Damerell, V. Swamy, H. Tamir, J. C. Taylor, R. E. Tennant, W. Thompson, A. Thompson, S. Tomasio, C. Tomlinson, I. S. Tsurupa, I. S. Tsurupa, A. Tumber, I. Vakonakis, R. P. van Rij, L. Vangeel, F. S. Varghese, M. Vaschetto, E. B. Vitner, V. Voelz, A. Volkamer, F. von Delft, A. von Delft, M. Walsh, W. Ward, C. Weatherall, S. Weiss, K. M. White, C. F. Wild, K. D. Witt, M. Wittmann, N. Wright, Y. Yahalom-Ronen, N. K. Yilmaz, D. Zaidmann, I. Zhang, H. Zidane, N. Zitzmann and S. N. Zvornicanin, *bioRxiv*, 2023, 2020.10.29.339317.

25 S. Ruiz-Carmona, D. Alvarez-Garcia, N. Foloppe, A. B. Garmendia-Doval, S. Juhos, P. Schmidtke, X. Barril, R. E. Hubbard and S. D. Morley, *PLoS Comput Biol*, , DOI:10.1371/JOURNAL.PCBI.1003571.

26 E. Resnick, A. Bradley, J. Gan, A. Douangamath, T. Krojer, R. Sethi, P. P. Geurink, A. Aimon, G. Amitai, D. Bellini, J. Bennett, M. Fairhead, O. Fedorov, R. Gabizon, J. Gan, J. Guo, A. Plotnikov, N. Reznik, G. F. Ruda, L. Díaz-Sáez, V. M. Straub, T. Szommer, S. Velupillai, D. Zaidman, Y. Zhang, A. R. Coker, C. G. Dowson, H. M. Barr, C. Wang, K. V. M. Huber, P. E. Brennan, H. Ovaa, F. von Delft and N. London, *J Am Chem Soc*, 2019, **141**, 8951–8968.

27 T. Mühlethaler, L. Milanos, J. A. Ortega, T. B. Blum, D. Gioia, B. Roy, A. E. Prota, A. Cavalli and M. O. Steinmetz, *Angew Chem Int Ed Engl*, , DOI:10.1002/ANIE.202204052.

28 T. Mühlethaler, L. Milanos, J. A. Ortega, T. B. Blum, D. Gioia, B. Roy, A. E. Prota, A. Cavalli and M. O. Steinmetz, *Angewandte Chemie International Edition*, 2022, **61**, e202204052.

29 B. Goldman, S. Kearnes, T. Kramer, P. Riley and W. P. Walters, *J Med Chem*, 2022, **65**, 7073–7087.

30 M. Schuller, G. J. Correy, S. Gahbauer, D. Fearon, T. Wu, R. E. Díaz, I. D. Young, L. C. Martins, D. H. Smith, U. Schulze-Gahmen, T. W. Owens, I. Deshpande, G. E. Merz, A. C. Thwin, J. T. Biel, J. K. Peters, M. Moritz, N. Herrera, H. T. Kratochvil, A. Aimon, J. M. Bennett, J. B. Neto, A. E. Cohen, A. Dias, A. Douangamath, L. Dunnett, O. Fedorov, M. P. Ferla, M. R. Fuchs, T. J. Gorrie-Stone, J. M. Holton, M. G. Johnson, T. Krojer, G. Meigs, A. J. Powell, J. G. M. Rack, V. L. Rangel, S. Russi, R. E. Skyner, C. A. Smith, A. S. Soares, J. L. Wierman, K. Zhu, P. O Brien, N. Jura, A. Ashworth, J. J. Irwin, M. C. Thompson, J. E. Gestwicki, F. Von Delft, B. K. Shoichet, J. S. Fraser and I. Ahel, *Sci Adv*, , DOI:10.1126/SCIADV.ABF8711.

31 C. N. Johnson, C. Adelinet, V. Berdini, L. Beke, P. Bonnet, D. Brehmer, F. Calo, J. E. Coyle, P. J. Day, M. Frederickson, E. J. E. Freyne, R. A. H. J. Gilissen, C. C. F. Hamlett, S. Howard, L. Meerpoel, L. Mevellec, R. McMenamin, E. Pasquier, S. Patel, D. C. Rees and J. T. M. Linders, *ACS Med Chem Lett*, 2015, **6**, 31–36.

32 B. D. Belviso, R. Caliandro, M. De Candia, G. Zaetta, G. Lopopolo, F. Incampo, M. Colucci and C. D. Altomare, *J Med Chem*, 2014, **57**, 8563–8575.

33 Y. Shi, D. Sitkoff, J. Zhang, H. E. Klei, K. Kish, E. C. K. Liu, K. S. Hartl, S. M. Seiler, M. Chang, C. Huang, S. Youssef, T. E. Steinbacher, W. A. Schumacher, N. Grazier, A. Pudzianowski, A. Apedo, L. Discenza, J. Yanchunas, P. D. Stein and K. S. Atwal, *J Med Chem*, 2008, **51**, 7541–7551.

34 Y. Patel, V. J. Gillet, T. Howe, J. Pastor, J. Oyarzabal and P. Willett, *J Med Chem*, 2008, **51**, 7552–7562.

35 S. Wills, R. Sanchez-Garcia, S. D. Roughley, A. Merritt, R. E. Hubbard, T. Dudgeon, J. Davidson, F. von Delft and C. M. Deane, *bioRxiv*, 2022, 2022.12.15.520559.

36 A. T. Pagnamenta, R. S. Belles, B. A. Salbert, I. M. Wentzensen, M. J. G. Sacoto, F. J. R. Santos, A. Caffo, M. Ferla, B. Banos-Pinero, K. Pawliczak, M. Makvand, H. Najmabadi, R. Maroofian, T. Lester, A. L. Yanez-Felix, C. E. Villarroel-Cortes, F. Xia, K. Al Fayez, A. Al Hashem, D. Shears, M. Irving, A. C. Offiah, A. Kariminejad and J. C. Taylor, *Clin Genet*, , DOI:10.1111/CGE.14324.