Fragmenstein: predicting protein-ligand structures of compounds derived from known crystallographic fragment hits using a strict conserved-binding–based methodology

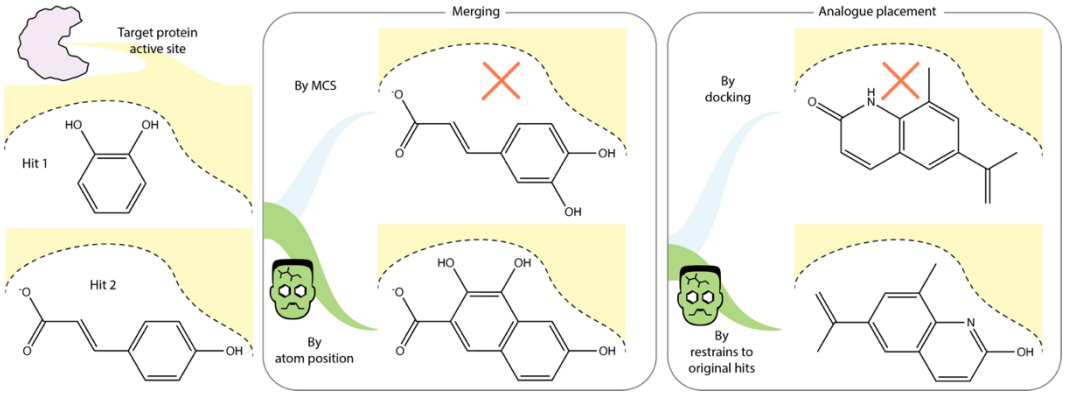
Matteo P. Ferla*[abc], Rubén Sánchez-García [a], Rachael E. Skyner [de], Stefan Gahbauer [f], Jenny C. Taylor [c], Frank von Delft [bdg], Brian D. Marsden [bd] and Charlotte M. Deane[a]

a. Oxford Protein Informatics Group, Department of Statistics, University of Oxford, UK
b. Centre for Medicine Discoveries, Nuffield Department of Medicine, University of Oxford, UK
c. NIHR Oxford BRC Genomic Medicine, Wellcome Centre for Human Genetics, University of Oxford, UK
d. Diamond Light Source, Science and Technology Facilities Council, UK
e. OMass Therapeutics, ARC Oxford, UK
f. Department of Pharmaceutical Chemistry, University of California San Francisco, San Francisco, USA
g. Department of Biochemistry, University of Johannesburg, South Africa

* matteo.ferla@stats.ox.ac.uk

Abstract

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 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 plausible compound. This method is successful in showing the importance of using the coordinates of known binders when predicting the conformation of derivative compounds 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 parent hits in a single round.



1. Introduction

1.1 Fragment-based drug discovery is a standard methodology in drug discovery that leverages the similar binding mode between analogues.

In the early stages of drug discovery fragment-based drug discovery (FBDD) has emerged as a standard methodology [1]. It 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 protein-ligand interaction information from these smaller proximal molecules to design larger derivative 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 [1–3].

Regardless of whether informative structural information is available for initial fragment hits, by far the most common strategy is to enumerate derivative 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 this approach negatively affect successfulness of the searches.

1.2 Docking approaches as conformational filters do not sufficiently leverage information from existing protein-ligand structures when predicting the conformation of derivative 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 [4]. Docking protocols find the energetic minimum according to the parameters of the force-field used to approximate the system, 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 reproduce roughly only half of all compounds docked to an RMSD of less than 2 Å [5]. An approach to improve this poor fidelity to the parent hits is by adding constraints to pharmacophores or to key atoms on the protein [6]. Another limitation stems from the fact most docking algorithms generate a set of small molecule conformers before docking which, especially for larger and more flexible small molecules, may all greatly diverge from the empirical crystallographic protein-bound conformer. Whereas it is straightforward to embed the conformer of a derivative compound with the conformation of a parent FBDD hits that is its direct substructure, it is non-trivial when the substructure overlaps are imperfect and between multiple hits, as will be addressed below.

1.3 Combinatorial approaches either disregard the position of hits or are unable to operate with overlapping hits.

When ligands are designed starting from fragment hits (rather than docking a subset of compounds in a dataset), the protein-ligand complex data available from initial fragment hit structures are often still not utilised until after initial enumeration. 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 [7] and AutoGrow4 [8]. Neither of these methods consider any 3D structural information from the protein or ligand in the initial enumeration step.

Some methods do consider some spatial information from the protein. DeLinker [9] 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 unaware of the protein environment around the ligands it is designing from. GANDI takes protein coordinates into consideration to filter out potential clashes [10], 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 [11]. 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 [12].

None of the methods discussed thus far consider the 3D conformation of overlapping hits. An algorithm that stands out in this respect is BREED [13], 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's a commercial product. In practice, in most situations, fragment merging is most often done by eye [14].

1.4 Fragmenstein generates energetically feasible protein-bound conformers that obey one or more parent hits

To address the above problems, we developed Fragmenstein. The governing idea behind Fragmenstein is striving for fidelity to the position of the inspiring hits based upon the assumption that the derivative compounds bind in a very similar way. The crucial difference is that the conformers are generated by stitching together the atoms of the parent hits for both de novo generation (combination) and for docking-like approach (placement), and subsequently minimised in place. To achieve this several tactics are employed to overcome certain issues, such as mapping partial overlaps to multiple parent molecules, merging rings and correcting impossible topologies.

2. Methods

2.1 Fragmenstein is a Python package with few dependencies.

The Fragmenstein codebase is a modular Python package that is dependent on RDKit [15] for compound manipulation, PyRosetta [16] for energy 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+. 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/oxpig/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. Fragmenstein has two main functionalities (Figure 1): fragment hit combination and derivative 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 parent hits and (ii) the energy minimisation of the compound under strong constraints.

The two functionalities can be used as a single continuous workflow (*viz.* GitHub repository). First, fragments are combined (merged/linked) with Fragmenstein, then a search is conducted via the SmallWorld server (https://sw.docking.org/) [17] for purchasable analogues in Enamine REAL or equivalent supplier (i.e. analogue-by-catalogue), and lastly candidate compounds are placed with Fragmenstein in order to be ranked.

The operations performed are described in the GitHub repository and outlined in Figure 1.

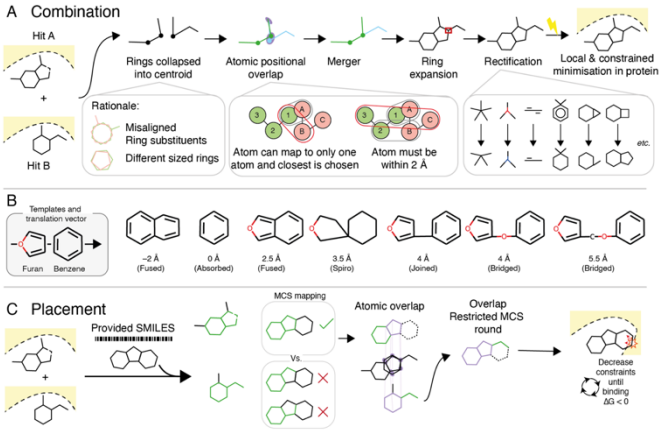


Figure 1. Combination and placement operations and their rules within Fragmenstein. A. Steps in a combination operation. For combinations, the positional overlap is calculated with any ring collapsed. This is done to prevent overlap issues (first inset, detail in SI Figure 1). Both rules share the atomic 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 benzene 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 stitched together conformer is created, it is energy minimised locally, with strong constraints and with a topology parameterised from an ideal conformer.

2.2 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) [18] and Mac1 domain of SARS-COV-2 NSP3 (macrodomain ADP-ribosylhydrolase) [19], were downloaded from Fragalysis (https://fragalysis.diamond.ac.uk/) [20] and filtered for inclusion in the DSi-Poised library [21]. 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 [16]. Their hits were combined (merged/linked) with the aim of quantifying 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[22]. These were placed with the PyRosetta framework of Fragmenstein (Igor). BREED [13] was also run with 1.5 Å cut-off and with the "untangle" setting disabled to increase number of compounds generated even if overly connected, but the limited results precluded its benchmarking. Interactions were determined with PLIP [23]. Interactive pages of results were created in MichelaNGLo [24].

2.3 MPro was used to assess the accuracy of placements of derivative compounds.

The information of which fragment hits were parents for which crystallised derivative compounds was taken from the Moonshot GitHub repository[18], but was reduced to contain only the relevant parent 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 derivative to not bias the selection (*cf,* code in repository). The common protein template used was PDB:6LU7, which was minimised as describe above. Fragmenstein was run with the tweak 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. Note that the latter functionality is automatic in Fragmenstein if the SMILES to be placed has a dummy atom (* in SMILES) or the warhead conversion code within Fragmenstein is called.

RDock was used as a benchmark for pharmacophore-constrained docking [25]. 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 parent 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.

3. Results

3.1 A retrospective placement of 100 compounds based on their parent reveals much strong agreement with the crystal structures than pharmacophoric constraints.

A key underlying hypothesis is the derivative compounds bind in a very similar way to their parent hits. Fragmenstein merges fragments by first stitching together the positioned atoms of the parent fragments prior and then locally minimising under strong constraints, without relying on previously generated conformers. To test this the dataset from the Covid Moonshot project was used as it contains a large panel of hit-inspired derivative compounds is available [18]. The Covid Moonshot project was a collaborative SAR-COV-2 protease fragment-based drug discovery project that relied on user submitted ideas of derivative compounds. 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 parent, 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

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. In a real-world scenario this calibration strategy is not possible since the crystallographic poses are not available, consequently, the results presented here are likely an overestimation of the actual performance.

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

First, to demonstrate the need for the thermodynamic corrections in the final step of Fragmenstein, the placement of a pair of derivative 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 parent 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.

Second, to demonstrate the use of user correction, the placement of the derivative 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 parent 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 (an equally plausible orientation given the electron density and required for the elaboration)

at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-moonshot).
The importance of exploiting the structural information of the parent hits is illustrated by the fact that 64% of the proposed merges were found to fully preserve the pose of their parent 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. 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 parent hits. Figure 2B shows that, while including constraints improves the docking performance, Fragmenstein still outperforms rDock, which was 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. The inability to sample a conformer that perfectly matches the crystallographic one underlies 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 parent hits.

3.2 On two datasets, Fragmenstein proposes 49 and 24 easily accessible derivative compounds from the combination of 34 and 44 parent hits.

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) and Mac1 (a nucleosyl-peptide hydrolase from SARS-COV-2) from previous fragment screens were chosen and the initial hits that originated from a library designed to facilitate synthetic derivatives (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 (ΔΔG < 0) without excessive deviation from the positions of the parent 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 [28]. Therefore, Fragmenstein suggests synthetically

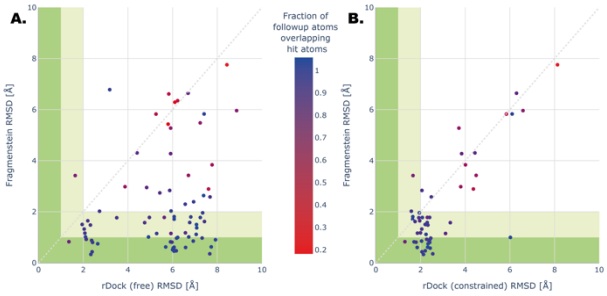


Figure 2. Accuracy of placement of COVID19 MPro1 Moonshot compounds. Derivative compounds in the COVID19 MPro1 Moonshot project which had a stated parent (manually adjusted) were placed with Fragmenstein and docked with rDock either freely or with pharmacophore constraints. The initial dataset contained 100 fragment-inspired compounds, but 23 were discarded (because the crystal structure of derivative had no overlapping atoms with the parent, the reactive derivative was non-covalent in the crystal structure and/or Fragmenstein failed to minimise the derivative 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 derivative to the hits due the convoluted overlay. Individual models can be investigated at https://michelanglo.sgc.ox.ac.uk/r/fragmenstein-moonshot.

accessible compounds that are predicted to follow the binding conformation of the parent fragment hits, which is an underpinning assumption in fragment-based drug discovery (*cf.* Figure 1).

3.2 The strict obedience to atomic positions by Fragmenstein is a strong filter whose effects may be misled by potentials and are unmasked when counting number of interactions.

As described above, a key point of Fragmenstein is obedience to parent hits. To emphasise the importance of fidelity to position of the initial hits of the initial hits of Mac1 were combined pairwise ignoring positional information in three different approaches. In one the parent hits were merged by maximum common substructure (MCS), in another by BRICS decomposition, and in a third with Fragmenstein but constrained to a single hit.
The minimisation of these in place via constraints to both the parent hits did not yield any acceptable poses, whereas 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.
This is because without the positional constraints the force-field dominates the placement by pushing towards a distant energy minimum. 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 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 contrasts with BRICS decomposition where the substructures of the parent 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 better than Fragmenstein (−0.25 kcal/mol/HA), whereas the median number of interactions was actually lower (0.17 interactions/HA).

	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.

3.3 Case examples

Fragmenstein can work 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 parent 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 parent 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 derivative atoms to map to specific atoms from the hit atoms in order to get the intended placement. An example of this is a parent 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 can apply to ligands [29].

An example of this is seen with tubulin inhibitor Todalam-4 [27]. 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.

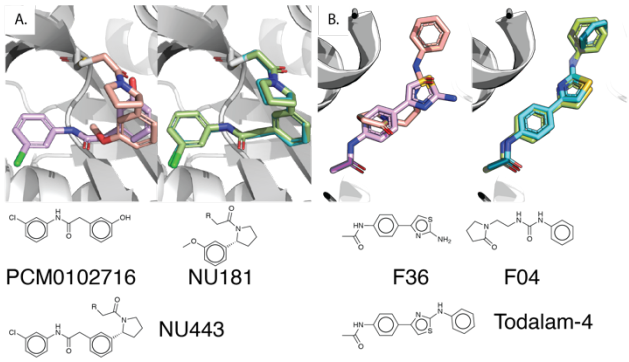


Figure 3. Retrospective Comparison of crystallised and placed derivative compound from NUDT7 study (A) and tubulin (B) study, illustrating an merger with hits that do not overlap cleanly and a merger requiring a user-defined mapping respectively.
In the NUDT7 study, the two hits NU181 (in lavender, LHS) 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 parent 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 (sky-blue: crystal, green: predicted). The aminothiazole ring is flipped between F04 and todalam-4 by design. A constructive observation of this derivative 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.

4. Discussion

4.1 Elaborations empirically follow their parent hits, so designs ought to do the same

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 parent hits, with the aim of being as faithful as possible to these without being energetically unfeasible.

Docking is often employed to shortlist compounds, however it 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 [3]. Were a docked derivative candidate to interact differently than its parents, the validity of its score would be rightfully put to question by an experimentalist. Several decomposition studies address the SAR additivity/superadditivity of certain functional groups[30–33] and how the binding mode is maintained crystallographically. Here, the inverse direction is taken and is found also to be consistent; in Figure 2 it was shown

that in the Covid-Moonshot dataset of the crystallised derivative compounds that bound similarly to their parent (69%), 82% are placed by Fragmenstein with an RMSD under 2 Å compared to 22% by pharmacophore-constrained docking. Confirming the importance of obeying the position of the atoms in the parent hits.
Fragmenstein has a very high success rate in combining parent hits and yields several virtual compounds in make-on-demand space (Table 1). Fragmenstein aims to preserve the interactions of the parent hits unlike other methods. Nevertheless, in assessing the elaborations, 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 there are fewer interactions.

4.2 A simple energy score for exploration is unsuitable for shortlisting virtual compounds for purchase or synthesis

Ranking virtual compounds from a screen via a predicted energy metric is less than ideal in general: a principle that also applies to Fragmenstein. With Fragmenstein in particular, the energy estimate is not of a global energy minimum, but a minimum highly constrained to the RMSD between the placed coordinates and parent hits: the RMSD should therefore be considered alongside the predicted potential.

In a pipeline, where fragment hits are combined, analogues identified by catalogue, and then placed, the next challenge becomes choosing which compounds to purchase, a problem shared with other methodologies. Three operations are commonly performed: filtering, sorting, and clustering. One possible filter is supplier driven, namely the removal of compounds above a given price point or with unworkable delivery times. Another possible filter is the wholesale removal of compounds with substructures that may cause assay interference, such as fluorescence or PAINS, or may be toxic (*e.g.* Ghose or REOS filters), or may not be drug-like (*eg.* Lipinski rules)[34, 35]. Whereas sorting by predicted energy or similar score is the simplest approach, it is less suitable in the real world than a combination of different metrics in addition to score or number of interactions. One factor is risk, whereas a conservative elaboration may be more likely to bind, more information may be gained from a riskier derivative compound. A variety of other factors could be considered such as a penalty for rotatable bonds on account of entropic loss from the decrease in degrees of rotational freedom upon binding. One further step, especially useful for hit discovery, is clustering by the interactions formed. These various steps together better reflect a drug discovery campaign as they allow a set of virtual compounds with desired properties and diverse binding modalities to be shortlisted as opposed to simply by predicted energy.

4.3 Fragmenstein can be paired with catalogue searches and decomposition.

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 [36–39]. A recent published approach, for example for fragment joining enumerate all purchasable compounds that contain substructure of pairs of hits and places these with Fragmenstein [36].
For close compounds, the torsion of the link may be highly constrained by the substructures from the parent hits, which is exactly the sort of problem Fragmenstein can address as demonstrated in its role in the identification of a IC_{50} 430 nM inhibitor against SARS-COV-2 Mac1[19, 40] (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 (an analogues-by-catalogue approach) [17], which can be placed by Fragmenstein. A similar approach was used in the SARS-COV-2 Mac1 study[19] (using Arthor, https://arthor.docking.org/[17]). 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 parents may yield a compound that is far removed from make-on-demand space (e.g. Supplementary figure 5, a clear planar 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.
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.
Beyond drug discovery, Fragmenstein has found uses in biochemistry settings by virtue of allowing the change of a crystallographically amenable analogue with the native substrate, *e.g.* the non-hydrolysable guanosine imidotriphosphate (GNP) for guanosine triphosphate (GTP) [41].

5. Conclusions

Fragmenstein is first and foremost a tool that strictly obeys the parent 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 chemist.

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**: Review & Editing. **JC. Taylor**: Supervision, Funding acquisition. **F. von Delft**: Conceptualization, Funding acquisition **BD. Marsden**: Supervision. **CM. Deane**: Supervision, Conceptualization, Validation, Writing - Review & Editing, Funding acquisition.

Conflicts of interest

There are no conflicts to declare.

Funding

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].

Acknowledgements

We would like to thank John Irwin for hosting sw.docking.org, and all the members of the Covid Moonshot consortium who designed compounds.

References

1. Erlanson DA. Introduction to fragment-based drug discovery. Top Curr Chem. 2012;317:1–32.
2. de Souza Neto LR, Moreira-Filho JT, Neves BJ, Maidana RLBR, Guimarães ACR, Furnham N, et al. In silico Strategies to Support Fragment-to-Lead Optimization in Drug Discovery. Frontiers in Chemistry. 2020;8:93.
3. Malhotra S, Karanicolas J. When Does Chemical Elaboration Induce a Ligand To Change Its Binding Mode? J Med Chem. 2017;60:128–45.
4. Torres PHM, Sodero ACR, Jofily P, Silva-Jr FP. Key topics in molecular docking for drug design. Int J Mol Sci. 2019;20:4574.
5. Houston DR, Walkinshaw MD. Consensus docking: Improving the reliability of docking in a virtual screening context. J Chem Inf Model. 2013;53:384–90.
6. Curran PR, Radoux CJ, Smilova MD, Sykes RA, Higueruelo AP, Bradley AR, et al. Hotspots API: A Python Package for the Detection of Small Molecule Binding Hotspots and Application to Structure-Based Drug Design. J Chem Inf Model. 2020;60:1911–6.
7. Liu T, Naderi M, Alvin C, Mukhopadhyay S, Brylinski M. Break Down in Order to Build Up: Decomposing Small Molecules for Fragment-Based Drug Design with eMolFrag. J Chem Inf Model. 2017;57:627–31.
8. Spiegel JO, Durrant JD. AutoGrow4: an open-source genetic algorithm for de novo drug design and lead optimization. Journal of Cheminformatics 2020 12:1. 2020;12:1–16.
9. Imrie F, Bradley AR, Van Der Schaar M, Deane CM. Deep Generative Models for 3D Linker Design. J Chem Inf Model. 2020;60:1983–95.
10. Dey F, Caflisch A. Fragment-based de Novo Ligand design by multiobjective evolutionary optimization. J Chem Inf Model. 2008;48:679–90.
11. Imrie F, Hadfield TE, Bradley AR, Deane CM. Deep generative design with 3D pharmacophoric constraints. Chem Sci. 2021;12:14577–89.
12. Hadfield TE, Imrie F, Merritt A, Birchall K, Deane CM. Incorporating Target-Specific Pharmacophoric Information into Deep Generative Models for Fragment Elaboration. J Chem Inf Model. 2022;62.
13. Pierce AC, Rao G, Bemis GW. BREED: Generating novel inhibitors through hybridization of known ligands. Application to CDK2, P38, and HIV protease. J Med Chem. 2004;47:2768–75.
14. Nikiforov PO, Surade S, Blaszczyk M, Delorme V, Brodin P, Baulard AR, et al. A fragment merging approach towards the development of small molecule inhibitors of Mycobacterium tuberculosis EthR for use as ethionamide boosters. Org Biomol Chem. 2016;14:2318–26.
15. Riniker S, Landrum GA. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. J Chem Inf Model. 2015;55:2562–74.
16. Chaudhury S, Lyskov S, Gray JJ. PyRosetta: a script-based interface for implementing molecular modeling algorithms using Rosetta. Bioinformatics. 2010;26:689.
17. Irwin JJ, Tang KG, Young J, Dandarchuluun C, Wong BR, Khurelbaatar M, et al. ZINC20 - A Free Ultralarge-Scale Chemical Database for Ligand Discovery. J Chem Inf Model. 2020;60:6065–73.
18. Boby ML, Fearon D, Ferla M, Filep M, Koekemoer L, Robinson MC, et al. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. Science. 2023;382:eabo7201.
19. Gahbauer S, Correy GJ, Schuller M, Ferla MP, Doruk YU, Rachman M, et al. Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. Proc Natl Acad Sci U S A. 2023;120:e2212931120.
20. Pearce NM, Skyner R, Krojer T. Experiences From Developing Software for Large X-Ray Crystallography-Driven Protein-Ligand Studies. Front Mol Biosci. 2022;9:861491.

21. Cox OB, Krojer T, Collins P, Monteiro O, Talon R, Bradley A, et al. A poised fragment library enables rapid synthetic expansion yielding the first reported inhibitors of PHIP(2), an atypical bromodomain. Chem Sci. 2016;7:2322–30.
22. Degen J, Wegscheid-Gerlach C, Zaliani A, Rarey M. On the art of compiling and using "drug-like" chemical fragment spaces. ChemMedChem. 2008;3:1503–7.
23. Salentin S, Schreiber S, Haupt VJ, Adasme MF, Schroeder M. PLIP: Fully automated protein-ligand interaction profiler. Nucleic Acids Res. 2015;43:W443–7.
24. Ferla MP, Pagnamenta AT, Damerell D, Taylor JC, Marsden BD. MichelaNglo: sculpting protein views on web pages without coding. Bioinformatics. 2020;36:3268–70.
25. Ruiz-Carmona S, Alvarez-Garcia D, Foloppe N, Garmendia-Doval AB, Juhos S, Schmidtke P, et al. rDock: a fast, versatile and open source program for docking ligands to proteins and nucleic acids. PLoS Comput Biol. 2014;10.
26. Resnick E, Bradley A, Gan J, Douangamath A, Krojer T, Sethi R, et al. Rapid Covalent-Probe Discovery by Electrophile-Fragment Screening. J Am Chem Soc. 2019;141:8951–68.
27. Mühlethaler T, Milanos L, Ortega JA, Blum TB, Gioia D, Roy B, et al. Rational Design of a Novel Tubulin Inhibitor with a Unique Mechanism of Action. Angew Chem Int Ed Engl. 2022;61.
28. Lee AA, Yang Q, Sresht V, Bolgar P, Hou X, Klug-Mcleod JL, et al. Molecular Transformer unifies reaction prediction and retrosynthesis across pharma chemical space. Chemical Communications. 2019;55:12152–5.
29. Drenth J, Mesters J. Principles of protein X-ray crystallography: Third edition. 2007.
30. Johnson CN, Adelinet C, Berdini V, Beke L, Bonnet P, Brehmer D, et al. Structure-based design of type II inhibitors applied to maternal embryonic leucine zipper kinase. ACS Med Chem Lett. 2015;6:31–6.
31. Belviso BD, Caliandro R, De Candia M, Zaetta G, Lopopolo G, Incampo F, et al. How a β-D-glucoside side chain enhances binding affinity to thrombin of inhibitors bearing 2-chlorothiophene as P1 moiety: Crystallography, fragment deconstruction study, and evaluation of antithrombotic properties. J Med Chem. 2014;57:8563–75.
32. Shi Y, Sitkoff D, Zhang J, Klei HE, Kish K, Liu ECK, et al. Design, structure-activity relationships, X-ray crystal structure, and energetic contributions of a critical P1 pharmacophore: 3-Chloroindole-7-yl- based factor Xa inhibitors. J Med Chem. 2008;51:7541–51.
33. Patel Y, Gillet VJ, Howe T, Pastor J, Oyarzabal J, Willett P. Assessment of additive/nonadditive effects in structure-activity relationships: Implications for iterative drug design. J Med Chem. 2008;51:7552–62.
34. Baell JB, Holloway GA. New substructure filters for removal of pan assay interference compounds (PAINS) from screening libraries and for their exclusion in bioassays. J Med Chem. 2010;53:2719–40.
35. Huggins DJ, Venkitaraman AR, Spring DR. Rational Methods for the Selection of Diverse Screening Compounds. ACS Chem Biol. 2011;6:208.
36. Wills S, Sanchez-Garcia R, Dudgeon T, Roughley SD, Merritt A, Hubbard RE, et al. Fragment Merging Using a Graph Database Samples Different Catalogue Space than Similarity Search. J Chem Inf Model. 2023;63:3423–37.
37. Zaidman D, Prilusky J, London N. ProsetTac: Rosetta based modeling of PROTAC mediated ternary complexes. J Chem Inf Model. 2020;60:4894–903.
38. Imrie F, Bradley AR, Van Der Schaar M, Deane CM. Deep Generative Models for 3D Linker Design. J Chem Inf Model. 2020;60.
39. Imrie F, Hadfield TE, Bradley AR, Deane CM. Deep generative design with 3D pharmacophoric constraints. Chem Sci. 2021;12.
40. Schuller M, Correy GJ, Gahbauer S, Fearon D, Wu T, Díaz RE, et al. Fragment binding to the Nsp3 macrodomain of SARS-CoV-2 identified through crystallographic screening and computational docking. Sci Adv. 2021;7.
41. Pagnamenta AT, Belles RS, Salbert BA, Wentzensen IM, Sacoto MJG, Santos FJR, et al. The prevalence and phenotypic range associated with biallelic PKDCC variants. Clin Genet. 2023. https://doi.org/10.1111/CGE.14324.