Active learning driven prioritisation of compounds from on-demand libraries targeting the SARS-CoV-2 main protease

Ben Cree,[a] Mateusz K. Bieniek, Siddique Amin, Akane Kawamura, and Daniel J. Cole*[1]

School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom

E-mail: daniel.cole@ncl.ac.uk

[a]B.C. and M.K.B. contributed equally to this work.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Abstract

FEgrow is an open-source software package for building congeneric series of compounds in protein binding pockets. For a given ligand core and receptor structure, it employs hybrid machine learning / molecular mechanics potential energy functions to optimise the bioactive conformers of supplied linkers and functional groups. Here, we introduce significant new functionality to automate, parallelise and accelerate the building and scoring of compound suggestions, such that it can be used for automated de novo design. We interface the workflow with active learning to improve the efficiency of searching the combinatorial space of possible linkers and functional groups, make use of interactions formed by crystallographic fragments in scoring compound designs, and introduce the option to seed the chemical space with molecules available from on-demand chemical libraries. As a test case, we target the main protease of SARS-CoV-2, identifying several small molecules with high similarity to molecules discovered by the COVID Moonshot effort, using only structural information from a fragment screen in a fully automated fashion. Finally, we order and test 19 compound designs, of which three show weak activity in a fluorescence-based Mpro assay, but work is needed to further optimise the prioritisation of compounds for purchase.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Introduction

Recent advances in structural biology, from sample preparation, to synchrotron infrastructure and data analysis pipelines, have transformed the throughput of protein-ligand complexes available to inform drug discovery campaigns.[1] When soaked with carefully designed compound libraries,[2] the numbers of small molecule (or fragment) structural hits can reach 10s or 100s against a single therapeutic target.[3] A frequently employed next step is to attempt to grow and/or link the hit compounds, using either custom synthesis[2] or ordering from catalogues of purchasable compounds.[4,5] However, chemical space is vast such that even choosing follow-up compounds for purchase from on-demand libraries, such as the readily accessible (REAL) Enamine database[6] ($>$ 5.5 bn compounds in 2022), becomes highly non-trivial.[7]

As such, attention is turning to cheminformatics and machine learning based algorithms for structure-based *de novo* hit expansion, linking and merging.[8] A wide range of approaches are available to build from initial structural biology data, including DeepFrag[9] that identifies promising fragments for addition to an input bound ligand, using a deep convolutional neural network, and DEVELOP[10] that combines 3D pharmacophoric constraints from the binding pocket with a graph-based deep generative model for R-group and linker design. The SILVR method enables an equivariant diffusion model to be conditioned to generate molecules based on a reference structure, such as a fragment from a crystallographic screen.[11] The V-SYNTHES approach makes use of on-demand libraries for hit-finding by decomposing compounds from purchasable databases into reactive scaffolds and synthons, and using the highest scoring docked fragments as seeds for further growth.[12] One particularly noteworthy example is the use of fragment merging to design hits against the nonstructural protein 3 (NSP3) of the severe acute respiratory syndrome-coronavirus-2 (SARS-CoV-2).[13] Fragments from a crystallographic screen were merged using the Fragmenstein package,[14] ensuring placement of molecular substructures onto the original fragments, and subsequently used as templates for searching on-demand chemical space. In this way, fragments were rapidly elaborated into a 0.4 μM hit (representing a >400-fold improvement in affinity).

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

While extremely promising, all of the above *de novo* design approaches suffer from some combination of the following issues: i) reliance on an approximate classical molecular mechanics force field or knowledge-based algorithm for generating and optimising binding poses, ii) use of an approximate objective function (usually a docking score) as a surrogate measure of binding affinity, iii) approximation of a rigid target receptor structure, and iv) limited synthetic tractability of the designed compounds. We therefore developed the FEgrow software as an open-source, interactive Jupyter notebook based workflow for building user-defined congeneric series of ligands in protein binding pockets to start to address some of these open questions (Figure 1A).[15] FEgrow grows user-defined functional groups (R-groups) off a constrained core of a known hit compound, thus incorporating input from structural biology and the expertise of the user in selecting synthetically tractable elaborations. Since publication, we have added functionality for connecting R-groups to the core via a flexible linker, which can be chosen from a library of those common to bioactive molecules.[16] In this way, users can choose from 1M+ combinations of linker and R-group from our distributed libraries (or upload their own R-group modifications). The modular workflow allows for the incorporation of state-of-the-art molecular modelling algorithms, such as the use of hybrid machine learning / molecular mechanics potential energy functions to optimise the ligand binding pose,[17,18] and the gnina convolutional neural network scoring function to predict the binding affinity.[19] We plan to expand the range of available optimisation algorithms and scoring functions as they become available (see Methods Section).

While interactive work is useful for small-scale studies, we have found it useful to automate the workflow for use on high performance computing (HPC) clusters, and since publication have added an application programming interface (API) to FEgrow (Figure 1B). This enables us to build virtual libraries with a common core, for example, using reaction-based generative scaffold decoration with LibInvent[21] or substructure searching of compound libraries,[22] and then rapidly build the compounds into the protein binding pocket with FEgrow. However, unless the libraries are designed using information from the binding pocket,

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

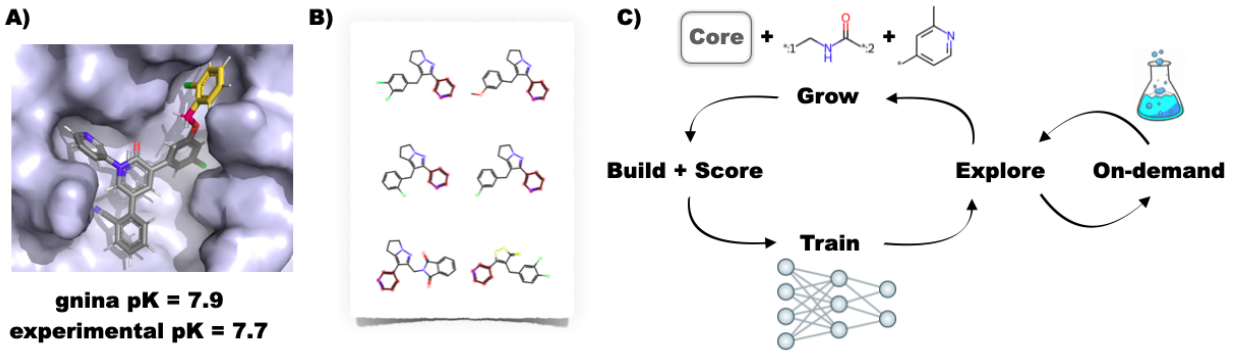

Figure 1: A) Example building and scoring of a SARS-CoV-2 inhibitor[20] using the interactive FEgrow workflow.[15] The fixed core (grey) is extended using a user-defined, flexible linker (pink) and R-group (yellow), and scored using gnina.[19] B) Compound libraries with substructures that match the rigid core can now be automatically grown and scored, treating the rest of the molecule as fully flexible. C) Proposed active learning cycle. Compounds are grown, built in the binding pocket and scored with FEgrow. The outputs are used to train a machine learning model, which is used to select the next batch of compounds. Optionally, the chemical space can be seeded using compounds available from on-demand chemical libraries.

time is wasted building and scoring compounds that are unlikely to be beneficial and it is still not feasible to routinely scan all possibilities.

Hence, rather than exhaustive or random searches of chemical space, we investigate here the use of active learning to elaborate compound design with FEgrow. The general idea behind this approach is that a subset of compounds is evaluated using an expensive design objective function (in this case the molecular growing and scoring algorithms in FEgrow) and used to train a machine learning model (Figure 1C).[23] The machine learning model then predicts the objective function for the remainder of the chemical space, and the next subset of molecules is picked for evaluation (for example, in order to optimise the objective or further explore the chemical space). By cycling through this procedure, the algorithm can iteratively make up for any lack of diversity in the initial training subset, and it is has been found previously that the most promising compounds can be identified by evaluating only a fraction of the total chemical space.

Several studies have investigated the effects of choices such as machine learning algorithm,

https://doi.org/10.26434/chemrxiv-2024-xczfb ORCID: https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

sample selection protocol and total dataset size on active learning efficiency for experimental and computational affinity predictions.[24–28] In general, active learning has been shown to increase enrichment of hits compared to either random or one-shot training of a machine learning model, at low additional cost, and to be relatively insensitive to choices of molecular representation, model hyperparameters and initial training subsets. Active learning has shown practical utility in prioritising compounds based on objective functions from docking[29–31] or free energy calculations.[25,26,32,33]

Here, we interface FEgrow with active learning to efficiently search the chemical space of linkers and R-groups from a user-defined vector. As well as using a docking score to guide optimisation, we also experiment with functions that combine other molecular properties, such as molecular weight, and 3D structural information, such as protein-ligand interaction profiles (PLIP).[34] To address the issue of synthetic tractability of the compound designs, we combine the workflow with regular searches of the Enamine REAL database to 'seed' the chemical search space with promising purchasable compounds. After testing and optimising the hyperparameters of the active learning models, we apply the algorithm to the prospective design of inhibitors of the main protease (MPro) of SARS-CoV-2, the virus responsible for the COVID-19 pandemic. This target has undergone extensive study in recent years. The COVID Moonshot Consortium used open science crowd-sourced designs, in combination with high-throughput structural biology and assays, free energy calculations, and machine learning driven synthetic route predictions, to generate a series of potent inhibitors.[4] Other notable approaches that include biological confirmation of hits have employed, for example, structure-based design starting from a drug repurposing study,[20] virtual screening of a curated collection of commercially available compounds,[35] a deep reinforcement learning model using pharmacophore and substructure matches with known inhibitors,[36] and a deep generative framework using only target sequence information as input (along with prioritisation based on factors such as docking score and retrosynthetic feasibility).[37] Here we employ active learning to prioritise compounds for purchase and testing from the Enamine

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

REAL database based only on early fragment hits. We suggest several novel designs that show activity in a fluorescence-based Mpro assay, as well as automatically generating several compounds that show high similarity to known Moonshot hits.

Methods

Workflow Design

The FEgrow software package is described in detail elsewhere.[15] Briefly, FEgrow aims to grow a ligand within a protein binding pocket, starting from a provided receptor structure, ligand core and growth vector (Figure 1A). Libraries comprising 2000 linkers[16] and around 500 R-groups, are provided, or users can supply their own. Merging is achieved using the RDKit package,[38] which also generates an ensemble of ligand conformations via the ETKDG algorithm,[39] with the atoms of the core strongly restrained to the input structure. That is, the default behaviour is to allow flexibility only in the regions of the grown linkers and R-groups. The ensemble of ligand structures is filtered to remove any that clash with the protein, and the remaining conformers are structurally optimised in the context of a rigid protein binding pocket using the OpenMM software.[18] During energy minimisation, the protein is treated using the AMBER FF14SB force field,[40] while intramolecular energetics of the ligand are described, where possible, using the ANI-2x machine learning potential.[17] Non-bonded interactions between the protein and ligand are described using a mechanical embedding scheme, that is, they use electrostatics and Lennard-Jones terms described by either the Open Force Field 'Sage'[41] or GAFF2[42] general force fields. The goal of this hybrid machine learning / molecular mechanics approach is to correct for known deficiencies in potential energy surfaces of classical force fields, while ensuring that optimisations are significantly faster than using full QM/MM.

The lowest energy structures are then output for scoring. In the first iteration of FEgrow, we used the gnina convolutional neural network (CNN), which has been jointly trained on

https://doi.org/10.26434/chemrxiv-2024-xczfb ORCID: https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

binding pose and affinity prediction.[19,43,44] We showed that the gnina 'CNNaffinity' scores (predicted pK) correlated reasonably well with experiment for ten series of congeneric inhibitors built using FEgrow.[15] Here, we add further options for scoring molecules based on protein-ligand interaction profile (PLIP),[34] molecular properties, or a combination thereof. For construction of the PLIP score, interactions formed in the available protein-fragment complex crystal structures were one-hot encoded to form a reference vector of desired interactions (here, hydrophobic, hydrogen-bonding, π-stacking, and salt bridge were all identified). A similar vector was constructed for the designed de novo compound, and its Tanimoto similarity to the reference vector used as the objective for optimisation. It has been argued that combining information from various properties can also be advantageous,[8] for example by using pharmacophore constraints in combination with docking scores, and we make use here of a simple, combined score (CS):

$$CS = \left(\frac{pK}{MW}\right) \times \left(\frac{PLIP}{0.3}\right) \times 100 \tag{1}$$

which aims to maximise the predicted gnina affinity (pK) and the protein-ligand interaction profile (PLIP) similarity to reference structures, while keeping the molecular weight (MW) low.

Active Learning

Active learning[23] is a subset of machine learning that is based on iteratively labeling data points from an unlabeled dataset (in our case, de novo compounds that are built into protein binding pockets and scored). The aim is to pick the most useful samples for training a surrogate model, whilst ultimately minimising the potentially expensive computation needed to find instances that maximise an objective function. There are two main components to an active learning workflow: the regression model, and the acquisition function. Every scored instance is used to train a specified machine learning model, with more examples refining

https://doi.org/10.26434/chemrxiv-2024-xczfb ORCID: https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

the model accuracy, which is then used to select new molecules to be built. In this work we consider and benchmark two models.

The first approach is gradient boosting machine (GBM), which is a random forest based technique, utilising ensembles of decision trees. These trees are created from random subsets of features (fingerprints), that are then used to make predictions. GBMs expand on traditional random forests by using the gradient of the error to construct trees specifically designed to minimise this error. Gradually increasing the number of relatively poor individual trees additively increases their predictive power (hence 'gradient boosted'). The second model is Gaussian Process (GP) regression, which is a Bayesian approach that makes predictions by assuming observations can be modelled by the probability distribution over possible reasonable (Gaussian) functions.[45] These Gaussian distributions are iteratively refined by the observation of new samples. Because model prediction is performed via a probability distribution, it natively incorporates uncertainty and other useful quantities, such as estimates of expected improvement of a given new sample.[46]

The acquisition function defines the method by which new molecules are picked at the start of each active learning cycle, with the simplest example being a 'greedy' approach, which directly selects the (currently predicted) highest scoring molecules. However, an acquisition function has to balance picking the best compounds, with the need to further refine the accuracy of the machine learning model. Picking the best scoring candidates in descending order might initially increase the objective function, but the algorithm will have the propensity to get stuck in local maxima and to be sensitive to the initial selection of training molecules.

There are a variety of alternatives that aim to avoid the problems of a simple greedy approach, and the approach used here is the upper confidence bound (UCB) uncertainty-based acquisition function.[47] UCB considers not just the value of the objective function, but also the variance of the prediction (model uncertainty), effectively biasing towards the selection of molecules about which the model is the least certain of the predicted score. The

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

UCB function is defined by:

$$UCB(x) = \mu(x) + \beta\sigma(x), \tag{2}$$

where $\mu(x)$ and $\sigma(x)$ are the mean and standard deviation of the regressor for molecule x, and β is a parameter controlling the degree of exploration (high β increases the chances that a molecule with moderate score but high uncertainty will be picked). The effects of the choice of machine learning model and acquisition function, as well as other active learning hyperparameters, are discussed later.

Database Search

A challenge for automated growing of linkers and R-groups, and for de novo design in general, is the synthetic tractability of the designed compounds. Approaches to address this limitation could include a synthetic accessibility score in the objective function[48] or the expert curation of libraries with known synthetic routes.[32] However, we wished to fully automate the design process, and be confident of acquiring compounds for rapid design-make-test-analyse cycles. We therefore make use of the rapidly-growing make-on-demand compound libraries as a surrogate measure of synthetic accessibility. Ideally, we might use the entire catalogue as a chemical space in which to perform the active learning. Although such an approach has been used as a one-off screen,[49] evaluating the regression models used here soon becomes prohibitively expensive in an active learning cycle. On the other hand, highly efficient methods have been developed for similarity and substructure searches of these libraries.[22] We therefore make use of these searches to seed the chemical space with compounds that are similar to the predicted actives at each step of the active learning cycle (Figure 1(C)). In this way, at the subsequent acquisition step, we enable the algorithm to pick compounds for growing and scoring that are likely to be scored highly (due to similarity with other highly scoring compounds) and available for purchase or synthesis (due to presence in on-demand libraries).

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

In detail, the Enamine REAL database of 4.5 B compounds was searched for similarity to designed molecules through the public interface to SmallWorld `https://sw.docking.org`, using a graph-edit-distance space search.[22] At each cycle, 100 new, top-scoring compounds were searched, and up to 100 of the most similar compounds from the REAL database were extracted per search query (using a maximum distance of 5 steps). This 10 K compound set was filtered for substructure match with the core using RDKit,[38] and those compounds that passed were added to the active learning search space. Active learning then selects compounds for scoring following Enamine enrichment, as usual, but there is no explicit bias to select compounds from the on-demand catalogue.

Computational Details

Protein input structures were taken from the set of noncovalent complexes crystallised early during the COVID-19 pandemic.[3] In particular, the input PDB: 5R83 was used as the receptor structure for active learning design, and Chimera was used to add hydrogen atoms.[50] The ligand was truncated to include only the pyridyl moiety, as this appeared in other available crystallised fragments in a consistent binding mode (PDB: 5RE4, 5REH, 5R84, 5RF3[3]) and with a suitable vector for growth into the binding pocket. The full set of 23 non-covalent complexes (that had ligands bound in areas of the pocket accessible by a growth vector) was additionally used for construction of the reference PLIP[34] interactions.

For testing of the active learning protocols, the chemical space was assembled by combining the pyridyl moiety with 508 R-groups[51] and 100 of the most common linkers[16] from the FEgrow library. A total of 47710 unique molecules were successfully grown into the binding pocket and scored using the gnina CNN scoring function.[19] A further 1656 molecules were assigned a penalty score of $pK = 0$ as they could not be embedded due to steric clash with the protein. In cases where rare errors occurred, such as a failure to assign force field parameters, the molecules were discarded completely.

The previously tested FEgrow molecule building protocol was applied throughout.[15] The

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

ETKDG algorithm[39] was used to generate 50 conformers, using a 0.5 Å root-mean-square similarity threshold. Any conformers with an atom closer than 1 Å to any atom in the protein was discarded. Energy minimisation was applied using a hybrid machine learning / molecular mechanics energy function in a mechanical embedding scheme.[15] The ANI-2x potential[17] was used for the ligand, in cases where all elements in the molecule are covered by the model, or the Open Force Field Sage[41] potential otherwise. The lowest energy conformer was retained for scoring.

An active learning library based on scikit[52] and modAL[53] python packages was adopted from another study.[26] A set of molecules to initialise the active learning cycle can be selected via RDKit's MaxMin picker[38] from the chemical space, or picked at random. The processing was parallelised using the python library Dask,[54] which supports a diverse set of technologies, including the Slurm Workload Manager that is deployed ubiquitously on high-performance computing clusters. Dask is used to secure resources (scheduling workers on Slurm), submitting work and retrieving results. The three major computationally-expensive components were parallelised: 1) building and scoring of the molecules, 2) computing the Morgan fingerprints, and 3) computing the Tanimoto similarity across the chemical space for the Gaussian Process modelling.

Results

Interfacing FEgrow with active learning enables efficient search of chemical space.

In order to investigate the performance of the active learning protocol, and the effect of machine learning hyperparameters, we built a labelled 'oracle' set of 47 K compounds using standard FEgrow input settings (see Computational Details). This is a larger set of compounds than would be typically built and scored against a target, but knowing the affinities of the full chemical space enables us to assess the performance of the active learning ap-

https://doi.org/10.26434/chemrxiv-2024-xczfb ORCID: https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

proach. The common core was selected to be a pyridyl fragment common to several early crystal structures of the SARS-CoV-2 main protease,[3] located in the S1 pocket with a vector pointing into the enzyme active site (Figure 2(a)).

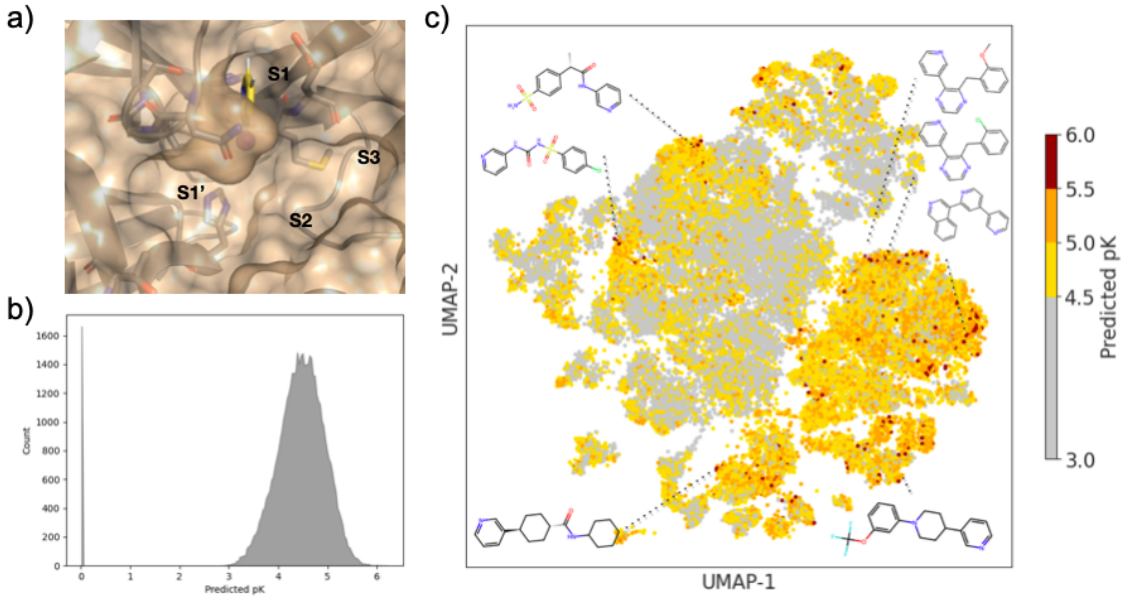

Figure 2: a) The position of the ligand core and definitions of binding pocket labels, the purple sphere is the hydrogen atom for replacement. b) Histogram of computed pK for the 47 K compound oracle dataset. c) UMAP of entire 47 K oracle chemical space, coloured by computed pK. 2D structures of representative strong binders are included.

Figure 2(b) shows the distribution of predicted binding affinities, computed using the gnina convolutional neural network scoring function[19] from FEgrow built structures. The scores are symmetrically distributed around pK = 4.5, with a maximum affinity of around 6.0, which is indicative of a set of low molecular weight (range between 100 and 350 Da, **Figure S1**), unoptimised compounds at the start of a hit finding effort. Indeed, it is at this stage where the options for expansion are vast, and strategies to suggest exploration of hits are particularly valuable. Note that compounds that could not be built (for example, due to steric clashes with the protein) are arbitrarily assigned a pK of zero, so that this information can be included in the active learning model.

Figure 2(c) further shows the UMAP projection of the chemical space, coloured by predicted pK. The visualisation shows a diverse composition of linkers and functional groups,

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

with well-spread clusters of the highest affinity binders, potentially providing a challenging search space for active learning. Figure 2(b) also shows locations in the chemical space of example linker and R-groups, attached to the pyridyl core, that make up the stronger predicted binders. Favourable predicted linkers include amides, sulfonylurea and various 6-membered ring heterocycles, and relatively bulky R-groups are feasible, which is generally expected given the size and shape of the binding pocket.[3,4] (Note that at this stage no consideration is given to synthetic accessibility or stability of the compound designs).

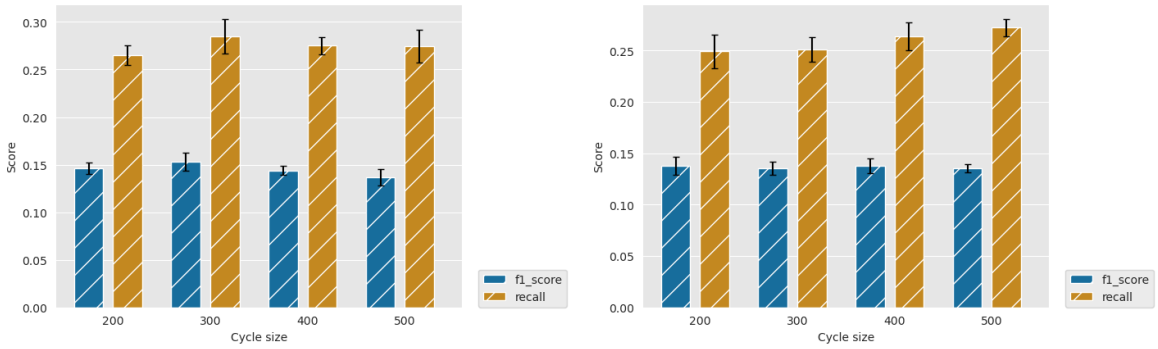

Figure 3: Recall and F1 score for diverse initial selection GBM (left) and GP (right) models, and greedy acquisition for identification of top 2 % scoring compounds for different cycle sizes. Error bars show standard errors over five runs.

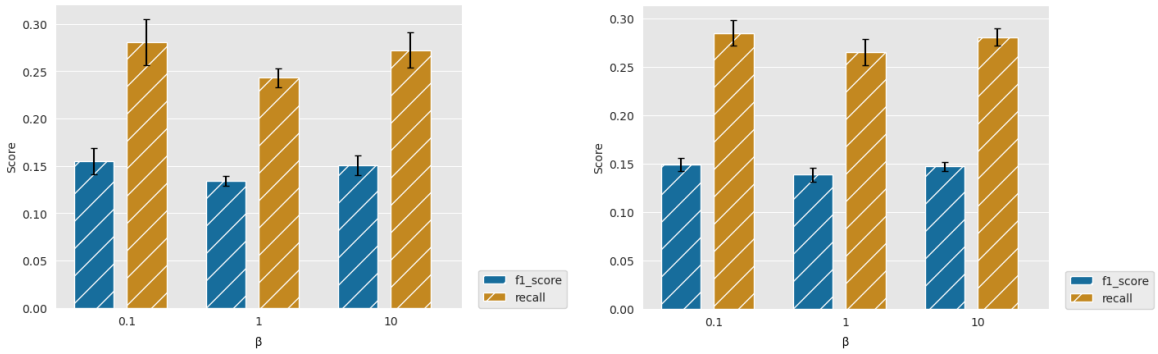

Figure 4: Recall and F1 score for diverse initial selection using GP and UCB acquisition (and varying β) with cycle sizes of 200 (left) and 400 (right) for identification of top 2 % scoring compounds. Error bars show standard errors over five runs.

We next sought to use active learning to accelerate the search through this chemical space, using the oracle to assess the performance of model hyperparameters, and using the predicted binding affinity as the optimisation target. In particular, we have investigated the

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

effects of initial compound selection (random or diverse), number of compounds picked per cycle (in the range 200–500), machine learning model (GBM or GP) and acquisition method (greedy or UCB). As discussed, the dependence of active learning efficiency on the choice of model parameters is well documented, and so we do not devote much space to it here.

By way of example, Figure 3 shows the effect of the number of compounds picked per cycle on model recall and precision (F1 score) for the two machine learning models (GBM and GP). For a fixed total number of compounds selected (here, 2500), one might expect the model to improve at small sample sizes (hence, more active learning cycles), but we find that the efficiency is already well converged when picking 500 per cycle. Similarly, the choice of machine learning model has little effect, with slightly higher metrics for the GBM model, but both recall and precision comparisons are within the error bars. Figure 4 further shows the effect of using the UCB uncertainty-based acquisition function, instead of greedy selection, in conjunction with the GP machine learning model. There is some small improvement in recall over greedy selection, but no significant change in the metrics used either as a function of cycle size or the β parameter in eq 2.

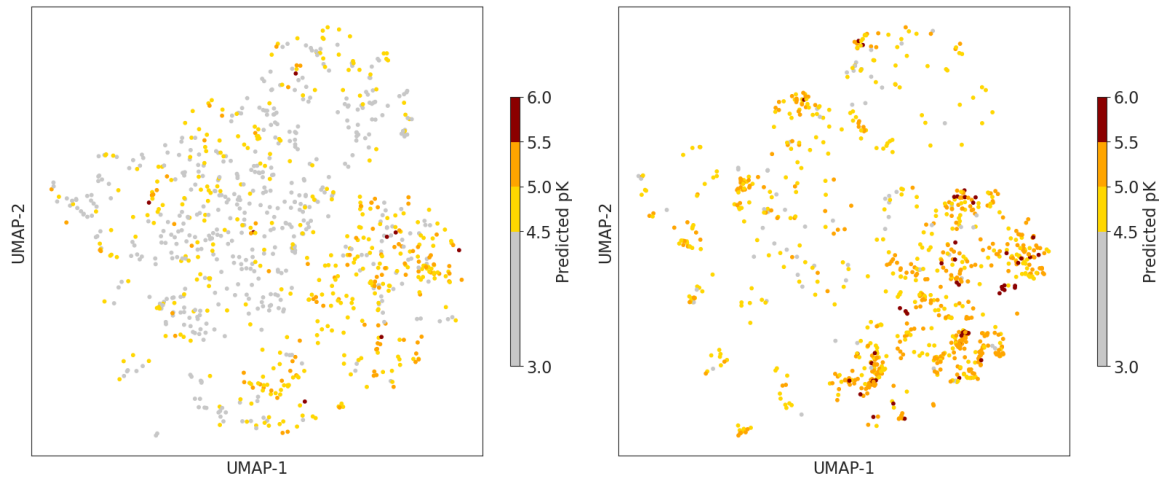

Figure 5: Difference in selection for first (left) and final (right) active learning cycles, showing a narrowing into areas predicted to be potent and avoiding unpromising areas.

Note that for the current dataset, random selection would give a recall of 0.05 and F1 score of 0.03 for identification of the top 2% of compounds. Therefore, with recall of around

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

0.25–0.30 for most of our experiments, we see efficiency improvements with active learning of around a factor of 5x compared to random selection. For reference, the growth and scoring of this compound set in FEgrow requires around 1000 cpuhrs, which is not prohibitive, but automated acceleration at no cost is clearly worthwhile.

In the next section, we choose to use a GP model with UCB acquisition function, with a cycle size of 200 and a diverse set of starting compounds. The overall accuracy of the chosen regression model (using $\beta = 10$), following training on 5 % of the dataset, is 0.97 pK units (**Figure S2**), which is competitive with typical models used in active learning with fingerprint-based representations.[27] Figure 5 shows a similar UMAP projection as in Figure 2, but now only showing compounds acquired by our chosen active learning model in the first (left) and final (right) cycles. We observe both a wide exploration of the chemical space, which is important to increase diversity in the final set, and a focusing of the explored regions in the final cycle to compounds with a higher predicted binding affinity, which is important for the use of the model to identify strong binders.

Active learning driven fragment expansion identifies potential SARS-CoV-2 MPro inhibitors.

Having established that the active learning protocols tested here are able to improve the efficiency of chemical space searches with FEgrow, we turn now to prospective design of potential noncovalent SARS-CoV-2 MPro inhibitors. A wealth of computational and experimental data has been generated for this target in recent years, but here we limit ourselves to structural information that was available in the early months of the COVID-19 pandemic. In particular, as in the previous section, we consider expansion of the pyridyl fragment (PDB: 5R83) along a vector into the binding pocket containing the catalytic cysteine (Cys145).[3] We now expand the size of the chemical space to an initial 250,000 molecules, built from the combination of supplied libraries of 500 linkers and 500 R-groups, such that full building and scoring of the space is prohibitively expensive for routine study. To address the issue of

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

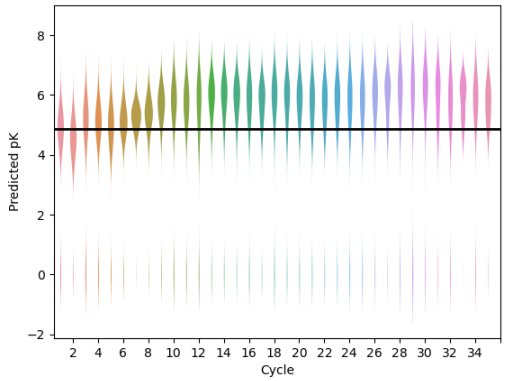

Figure 6: Active learning drives improvements in predicted binding affinity. A GP model is used, with UCB acquisition function ($\beta = 0.1$), a cycle size of 200 and a diverse set of starting compounds. The solid horizontal line shows the average score for 377 compounds randomly selected from the REAL database that were built with FEgrow.

synthetic feasibility of the output designs, we add an additional step in the active learning cycle (Figure 1C), whereby the chemical space is periodically seeded with compounds from the REAL database that are similar to the highest scoring compounds (see Methods). **Figure S3** demonstrates successful incorporation of the Enamine compounds into the active learning cycles, with a significant fraction of the built and scored compounds originating from this source.

Figure 6 shows an example design run, optimising the compounds for predicted pK using the gnina scoring function (further examples are given in the **Supporting Information**). The distribution of predicted affinity increases over the first 10 active learning cycles then starts to saturate with a mean predicted pK close to 6 (micromolar affinity). Over the full run, 95% of the compounds were successfully built (assigned $pK > 0$) and 15% had a predicted $pK > 6$. For comparison, we also extracted 1000 molecules at random that contained the pyridyl substructure from the REAL database used to seed the active learning cycles. For this set, 377 molecules (38%) could be successfully built, with an average predicted $pK = 4.9$ and only two compounds with predicted $pK > 6.0$ (0.2%).

Figure 7a) shows the highest scoring compound from this run, with a predicted affinity of 88 nM. The compound extends hydrophobic contacts into the S3 and S1' pockets, for ex-

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

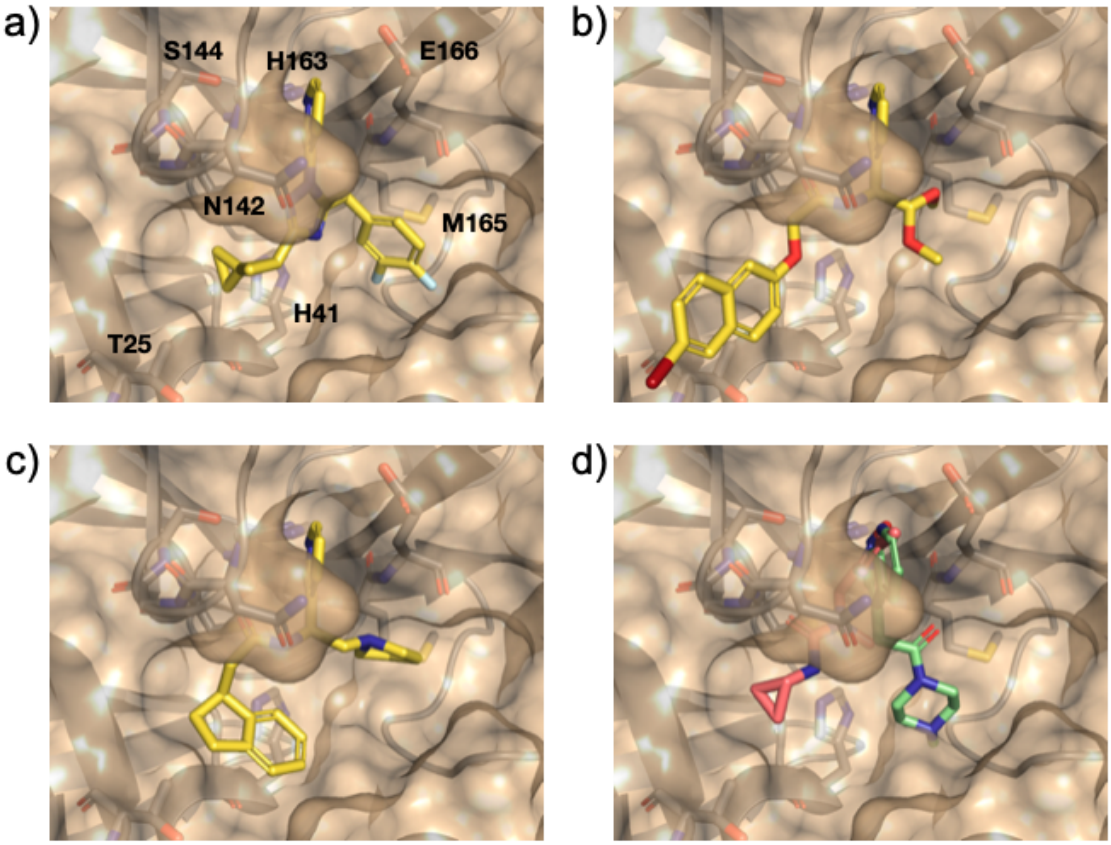

Figure 7: a) Top-scoring compounds optimised for a) predicted pK, b) protein-ligand interaction profile and c) combined scoring function. d) Fragment 5RGI shown in pink (H-bond donation by Gly143, Ser144, Cys145 and His163), and 5RF7 in green (hydrophobic and H-bond donation with Glu166).

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

ample with Met165 and Thr25, but despite this does not form any specific polar interactions (other than the original core interaction with His163). Since an early fragment screen had provided valuable information about the nature of potential protein–ligand interactions in this binding pocket, we sought to reduce the reliance on the gnina scoring function and drive the active learning towards compounds that recovered known crystallographic information (see Methods). Figure 7b) shows the top-scoring compound, as defined by the Tanimoto similarity to the vector of reference interactions. In this case, the grown molecule forms additional hydrogen bonding interactions with Asn142, Gly143, Ser144, Cys145 and Glu166, and hydrophobic interactions with Thr25 and Glu166. The majority of these interactions are recapitulated by, for example, fragments PDB: 5RGI and 5RF7 (Figure 7d)).

Finally, we sought to combine the strengths of both docking scores and crystallographic information to optimise a combined scoring function. Figure 7c) shows the top-scoring compound as defined by eq 1 after 33 cycles of active learning. Although this compound is scored much lower by the gnina scoring function (predicted affinity 2 μM), it extends into the S3 and S1' pockets and retains many of the interactions observed in Figure 7b) (e.g. hydrogen bonding interactions with Asn142, Gly143, Ser144, Cys145 and Glu166).

Analysis of hit compounds.

The top 500 compounds from each of four active learning runs (two optimising predicted pK, one optimising protein-ligand interactions, and one optimising the combined scoring function) were checked for availability from the Enamine store. Interestingly, very few of the top scored by predicted pK were available (four in total). This is likely due to an important unavailable building block(s), and could be mitigated in future by increasing diversity and/or including direct store queries in the search process. In any case, we focussed here on outputs from the remaining two runs, and submitted the top 10 protein-ligand interaction and top 25 combination scoring compounds for costing. Finally, a total of 19 designed compounds were purchased (of which 15 had been optimised used the combination score) based on

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

quoted price and excluding similar compounds (based on visual inspection). Two control compounds were also included; one known binder from a crystallographic fragment screen (Enamine ID: Z44592329; PDB: 5R83)[3] and one elaborated compound from the COVID Moonshot study (Enamine ID: Z4943052515 (literature IC_{50} 0.288 μM)).[4] The twenty one purchased compounds (**Figure S11**) were evaluated in a fluorescence-based Mpro activity assay at 1000, 500, 10 μM (**Figure S12**). Compounds **5** and **6** were excluded from the study due to solubility issues at 1000 μM in the assay conditions. Five compounds (**8**, **10**, **12**, **14** and **21** (the positive control[4])) showed reduction of Mpro activity $\leq$ 50% at 1000 μM. The IC_{50} values of these compounds, except **8** which displayed background autofluorescence, were further determined (Figure 8). Compounds **10**, **12** and **14** showed a concentration-dependent inhibition of Mpro activity (measured pIC_{50} 2.10, 3.01, 2.80 respectively). Nirmatrelvir, an orally bioavailable antiviral drug targeting Mpro, showed inhibition (pIC_{50} 6.01), which was slightly higher than the reported IC_{50} (0.022 μM[55]), likely due to the limit of the assay (the enzyme concentration was at 0.2 μM). Figure 9 shows the predicted structures of compounds **12** and **14** from the active learning design runs. Both compounds form hydrogen bonding interactions with the backbone of Glu166, as well as hydrophobic interactions in the S1' pocket.

Finally, to investigate whether the relatively low affinity of designed compounds is due to insufficient exploration of chemical space or the empirical objective functions used to optimise molecules, we performed a retrospective analysis of the designed compound space against known binders resulting from the COVID Moonshot crowd-sourced discovery campaign.[4] In particular, Figure 10 shows the three most similar compounds from the active learning runs (as defined by Tanimoto similarity search between RDKit Morgan fingerprints with a radius of 3 and size of 2048) to a curated set of 292 hit compounds. Considering that our FEgrow runs took as input only a single PDB receptor structure and pyridyl fragment core, it is clear that this fragment growing and on-demand library screening approach holds promise for suggesting biologically active compounds early in hit discovery campaigns. However,

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

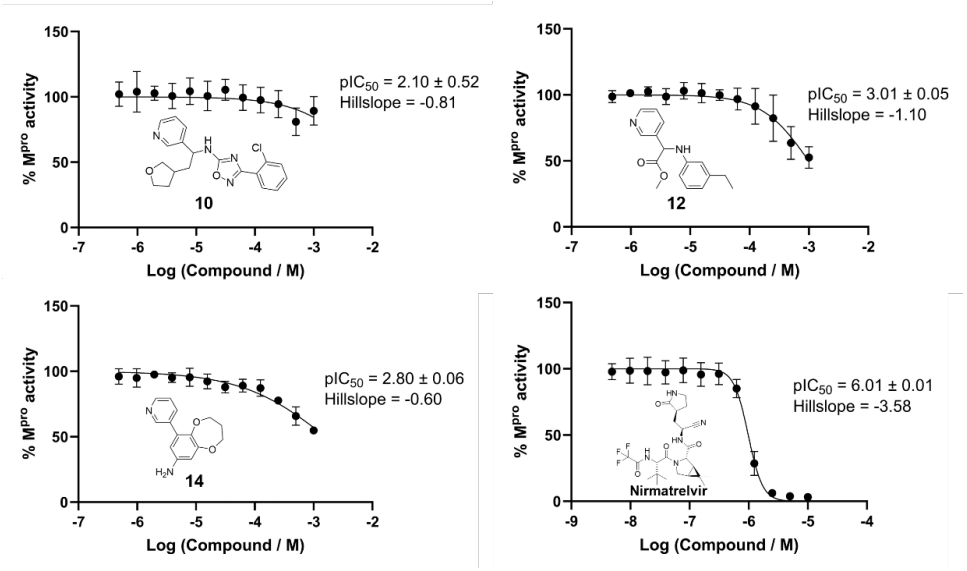

Figure 8: IC_{50} determination of selected compounds with Mpro. Compounds **10**, **12** and **14** were tested at a top concentration of 1000 μM. Nirmatrelvir was tested at a top concentration of 10 μM as a positive control. Datapoints presented as mean ± SD; pIC_{50} presented as mean ± SEM; two biological repeats consisting of three technical replicates. **10** consists of one biological repeat with three technical replicates. Conditions: Mpro (0.2 μM), 12-hour pre-incubation with compounds, 20 μM fluorescent substrate, 50 mM Tris-HCl (pH 7.3), 1 mM EDTA and temperature 25°C.

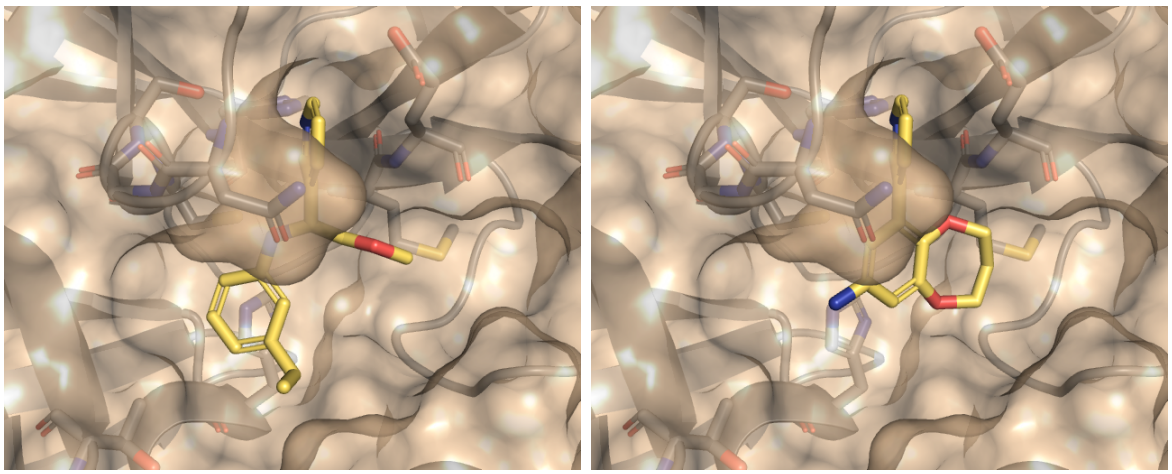

Figure 9: Predicted bound structures of compounds **12** (Z1470573089) and **14** (Z8969017446).

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

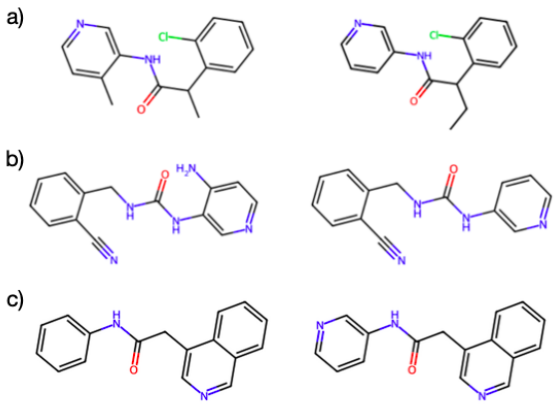

Figure 10: a) Experimental Moonshot compound (literature IC_{50} 17 μM) and most similar compound from this study, from active learning optimisation of predicted pK (β=10), b) Experimental Moonshot compound (literature IC_{50} 54 μM) and most similar compound from this study, from active learning optimisation of predicted pK (β=10), c) Experimental Moonshot compound (literature IC_{50} 57 μM) and most similar compound from this study, from active learning optimisation of combination scoring function.

401 further work is needed to ensure that the most promising compounds are located at the top
402 of ranked lists for synthetic prioritisation and testing.

403 Discussion and Conclusions

404 In this study, we have combined the FEgrow software, an open modular workflow for building
405 and scoring ligands in protein binding pockets, with active learning to guide and automate
406 chemical space searches for promising binders. In agreement with numerous other studies,[27]
407 we have shown that search efficiency is not too dependent on the hyperparameters of the
408 active learning model, which include the choice of regression model, the acquisition function
409 and number of compounds picked per cycle. For this particular study, we find efficiency
410 improvements of a factor of around 5x over random selection, which will aid throughput of
411 future prospective design efforts.

412 With the design of FEgrow, we hope to overcome some of the current limitations of de
413 novo drug design discussed in the Introduction. Some of these limitations are addressed in
414 the current study, and some will be addressed in future aided by ongoing advances in molec-

https://doi.org/10.26434/chemrxiv-2024-xczfb ORCID: https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0

ular modelling and machine learning. For example, we tackle the question of binding pose optimisation by using a fast and accurate machine learning potential (ANI-2x[17]) to describe the ligand energetics in a mechanical embedding scheme. However, with the flexibility of the FEgrow interface with OpenMM,[18] new models could be substituted in, and these are now approaching sufficient speed and accuracy (including for long-ranged interactions) such that the entire protein-ligand complex could be described using a single, consistent machine learning potential.[56,57] In this study, we made the approximation that the protein binding pocket is rigid and used a single receptor structure for design. However, now that ligand building and scoring is fully automated, future studies could use, for example, ensembles of receptor structures, which may be beneficial in cases where the pocket is more flexible.

A limitation of this and other similar studies is the choice of objective function in the active learning cycles. To demonstrate the flexibility of the FEgrow package, we demonstrated four design cycles here, two optimising for predicted affinity using the gnina CNN scoring function and two including a more direct optimisation of protein-ligand contacts extracted from crystallographic fragment screens. While we do not have enough data to assess the relative merits of these scoring functions, we expect the latter to be useful where experimental structural data exists, at least as part of a multi-objective optimisation in future.[58] As a flexible alternative to PLIP scores trained on system-dependent crystal structures, it has also been shown that transferable neural networks can be trained on the PDBbind structural database to recognise favourable protein-ligand interactions.[59]

As shown in Figure 1c), to address the issue of synthetic tractability of the de novo built compounds, we inserted regular queries of the Enamine REAL database into the active learning cycles. In this way, we can use the initial chemical space to train the active learning regression models, and then over time seed the chemical space with compounds that are both similar to predicted actives and purchasable. In this way, we were able to test the predictions of the active learning workflow with a turn around time of a few weeks from order to biological testing. Of the 19 designed compounds that were purchased here, three showed measurable

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

activity, but none approached the desired levels for further progression. Nevertheless, a similarity search showed the presence of effective inhibitors in the built chemical space, and so further investigation will focus on ranking compound designs ahead of purchase, perhaps via an extra stage of physics-based free energy calculations.[26]

Code Availability

FEgrow is freely available, with a set of tutorials, at `https://github.com/cole-group/FEgrow`.

Author Contributions

Ben Cree: Conceptualisation, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing - original draft.

Mateusz Bieniek: Conceptualisation, Data curation, Formal Analysis, Investigation, Methodology, Software, Validation, Visualisation, Writing - original draft.

Siddique Amin: Data curation, Formal Analysis, Investigation, Methodology, Writing - original draft.

Akane Kawamura: Resources, Supervision, Writing – review & editing.

Daniel Cole: Conceptualisation, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing – original draft, Writing – review & editing.

Competing Interests Disclosure

The authors declare no competing interests.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

Acknowledgement

D.J.C. and M.K.B. acknowledge support from a UKRI Future Leaders Fellowship (grant MR/T019654/1). B.C. and S.A are grateful for support from the EPSRC Centre for Doctoral Training in Molecular Sciences for Medicine (grant EP/S022791/1). This research made use of the Rocket High Performance Computing service at Newcastle University. We are grateful to Dr Mathew Martin, Dr Julia Hubbard, and Dr Sara Pintar for provision of the Mpro expression construct and to Prof. Christopher Schofield for the provision of Nirmatrelvir.

Supporting Information Available

References

(1) Douangamath, A.; Powell, A.; Fearon, D.; Collins, P. M.; Talon, R.; Krojer, R., T.and Skyner; Brandao-Neto, J.; Dunnett, L.; Dias, A.; Aimon, A.; Pearce, N. M.; Wild, C.; Gorrie-Stone, T.; von Delft, F. Achieving Efficient Fragment Screening at XChem Facility at Diamond Light Source. *J. Vis. Exp.* **2021**, *171*, e62414.

(2) Wood, D. J.; Lopez-Fernandez, J. D.; Knight, L. E.; Al-Khawaldeh, I.; Gai, C.; Lin, S.; Martin, M. P.; Miller, D. C.; Cano, C.; Endicott, J. A.; Hardcastle, I. R.; Noble, M. E. M.; Waring, M. J. FragLites—Minimal, Halogenated Fragments Displaying Pharmacophore Doublets. An Efficient Approach to Druggability Assessment and Hit Generation. *Journal of Medicinal Chemistry* **2019**, *62*, 3741–3752.

(3) Douangamath, A. et al. Crystallographic and electrophilic fragment screening of the SARS-CoV-2 main protease. *Nature Communications* **2020**, *11*, 5047.

(4) Boby, M. L. et al. Open science discovery of potent noncovalent SARS-CoV-2 main protease inhibitors. *Science* **2023**, *382*, eabo7201.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

(5) Mackinnon, S. R.; Krojer, T.; Foster, W. R.; Diaz-Saez, L.; Tang, M.; Huber, K. V. M.; von Delft, F.; Lai, K.; Brennan, P. E.; Arruda Bezerra, G.; Yue, W. W. Fragment Screening Reveals Starting Points for Rational Design of Galactokinase 1 Inhibitors to Treat Classic Galactosemia. *ACS Chemical Biology* **2021**, *16*, 586–595.

(6) Grygorenko, O. O.; Radchenko, D. S.; Dziuba, I.; Chuprina, A.; Gubina, K. E.; Moroz, Y. S. Generating Multibillion Chemical Space of Readily Accessible Screening Compounds. *iScience* **2020**, *23*, 101681.

(7) Warr, W. A.; Nicklaus, M. C.; Nicolaou, C. A.; Rarey, M. Exploration of Ultralarge Compound Collections for Drug Discovery. *Journal of Chemical Information and Modeling* **2022**, *62*, 2021–2034.

(8) Kuan, J.; Radaeva, M.; Avenido, A.; Cherkasov, A.; Gentile, F. Keeping pace with the explosive growth of chemical libraries with structure-based virtual screening. *WIREs Computational Molecular Science* **2023**, *13*, e1678.

(9) Green, H.; Koes, D. R.; Durrant, J. D. DeepFrag: a deep convolutional neural network for fragment-based lead optimization. *Chem. Sci.* **2021**, *12*, 8036–8047.

(10) Imrie, F.; Hadfield, T. E.; Bradley, A. R.; Deane, C. M. Deep generative design with 3D pharmacophoric constraints. *Chem. Sci.* **2021**, *12*, 14577–14589.

(11) Runcie, N. T.; Mey, A. S. SILVR: Guided Diffusion for Molecule Generation. *Journal of Chemical Information and Modeling* **2023**, *63*, 5996–6005.

(12) Sadybekov, A. A. et al. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. *Nature* **2022**, *601*, 452–459.

(13) Gahbauer, S. et al. Iterative computational design and crystallographic screening identifies potent inhibitors targeting the Nsp3 macrodomain of SARS-CoV-2. *Proceedings of the National Academy of Sciences* **2023**, *120*, e2212931120.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

(14) Ferla, M.; Sánchez-García, R.; Skyner, R.; Gahbauer, S.; Taylor, J.; von Delft, F.; Marsden, B.; Deane, C. Fragmenstein: predicting protein-ligand structures of compounds derived from known crystallographic fragment hits using a strict conserved-binding–based methodology. *ChemRxiv* **2024**,

(15) Bieniek, M.; Cree, B.; Pirie, R.; Horton, J.; Tatum, N.; Cole, D. An open-source molecular builder and free energy preparation workflow. *Commun. Chem.* **2022**, *5*, 136.

(16) Ertl, P.; Altmann, E.; Racine, S. The most common linkers in bioactive molecules and their bioisosteric replacement network. *Bioorganic & Medicinal Chemistry* **2023**, *81*, 117194.

(17) Devereux, C.; Smith, J. S.; Huddleston, K. K.; Barros, K.; Zubatyuk, R.; Isayev, O.; Roitberg, A. E. Extending the Applicability of the ANI Deep Learning Molecular Potential to Sulfur and Halogens. *J. Chem. Theory Comput.* **2020**, *16*, 4192–4202.

(18) Eastman, P. et al. OpenMM 8: Molecular Dynamics Simulation with Machine Learning Potentials. *The Journal of Physical Chemistry B* **2024**, *128*, 109–116.

(19) McNutt, A. T.; Francoeur, P.; Aggarwal, R.; Masuda, T.; Meli, R.; Ragoza, M.; Sunseri, J.; Koes, D. R. GNINA 1.0: molecular docking with deep learning. *J. Cheminf.* **2021**, *13*, 1–20.

(20) Zhang, C.-H. et al. Potent Noncovalent Inhibitors of the Main Protease of SARS-CoV-2 from Molecular Sculpting of the Drug Perampanel Guided by Free Energy Perturbation Calculations. *ACS Cent. Sci.* **2021**, *7*, 467–475.

(21) Fialková, V.; Zhao, J.; Papadopoulos, K.; Engkvist, O.; Bjerrum, E. J.; Kogej, T.; Patronov, A. LibINVENT: Reaction-based Generative Scaffold Decoration for in Silico Library Design. *Journal of Chemical Information and Modeling* **2022**, *62*, 2046–2063.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

(22) Irwin, J. J.; Tang, K. G.; Young, J.; Dandarchuluun, C.; Wong, B. R.; Khurelbaatar, M.; Moroz, Y. S.; Mayfield, J.; Sayle, R. A. ZINC20—A Free Ultralarge-Scale Chemical Database for Ligand Discovery. *Journal of Chemical Information and Modeling* **2020**, *60*, 6065–6073.

(23) Yu, J.; Li, X.; Zheng, M. Current status of active learning for drug discovery. *Artificial Intelligence in the Life Sciences* **2021**, *1*, 100023.

(24) Graff, D. E.; Shakhnovich, E. I.; Coley, C. W. Accelerating high-throughput virtual screening through molecular pool-based active learning. *Chem. Sci.* **2021**, *12*, 7866–7881.

(25) Khalak, Y.; Tresadern, G.; Hahn, D. F.; de Groot, B. L.; Gapsys, V. Chemical Space Exploration with Active Learning and Alchemical Free Energies. *Journal of Chemical Theory and Computation* **2022**, *18*, 6259–6270.

(26) Thompson, J.; Walters, W. P.; Feng, J. A.; Pabon, N. A.; Xu, H.; Maser, M.; Goldman, B. B.; Moustakas, D.; Schmidt, M.; York, F. Optimizing active learning for free energy calculations. *Artificial Intelligence in the Life Sciences* **2022**, *2*, 100050.

(27) Gorantla, R.; Kubincová, A.; Suutari, B.; Cossins, B. P.; Mey, A. S. J. S. Benchmarking Active Learning Protocols for Ligand-Binding Affinity Prediction. *Journal of Chemical Information and Modeling* **2024**, *64*, 1955–1965.

(28) van Tilborg, D.; Grisoni, F. Traversing Chemical Space with Active Deep Learning: A Computational Framework for Low-data Drug Discovery. *ChemRxiv* **2024**,

(29) Gentile, F.; Agrawal, V.; Hsing, M.; Ton, A.-T.; Ban, F.; Norinder, U.; Gleave, M. E.; Cherkasov, A. Deep Docking: A Deep Learning Platform for Augmentation of Structure Based Drug Discovery. *ACS Central Science* **2020**, *6*, 939–949.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

(30) Yang, Y.; Yao, K.; Repasky, M. P.; Leswing, K.; Abel, R.; Shoichet, B. K.; Jerome, S. V. Efficient Exploration of Chemical Space with Docking and Deep Learning. *Journal of Chemical Theory and Computation* **2021**, *17*, 7106–7119.

(31) Sivula, T.; Yetukuri, L.; Kalliokoski, T.; Käsnänen, H.; Poso, A.; Pöhner, I. Machine Learning-Boosted Docking Enables the Efficient Structure-Based Virtual Screening of Giga-Scale Enumerated Chemical Libraries. *Journal of Chemical Information and Modeling* **2023**, *63*, 5773–5783.

(32) Konze, K. D.; Bos, P. H.; Dahlgren, M. K.; Leswing, K.; Tubert-Brohman, I.; Bortolato, A.; Robbason, B.; Abel, R.; Bhat, S. Reaction-Based Enumeration, Active Learning, and Free Energy Calculations To Rapidly Explore Synthetically Tractable Chemical Space and Optimize Potency of Cyclin-Dependent Kinase 2 Inhibitors. *Journal of Chemical Information and Modeling* **2019**, *59*, 3782–3793.

(33) Gusev, F.; Gutkin, E.; Kurnikova, M. G.; Isayev, O. Active Learning Guided Drug Design Lead Optimization Based on Relative Binding Free Energy Modeling. *Journal of Chemical Information and Modeling* **2023**, *63*, 583–594.

(34) Adasme, M. F.; Linnemann, K. L.; Bolz, S. N.; Kaiser, F.; Salentin, S.; Haupt, V. J.; Schroeder, M. PLIP 2021: expanding the scope of the protein–ligand interaction profiler to DNA and RNA. *Nucleic Acids Research* **2021**, *49*, W530–W534.

(35) Glaab, E.; Manoharan, G. B.; Abankwa, D. Pharmacophore Model for SARS-CoV-2 3CLpro Small-Molecule Inhibitors and in Vitro Experimental Validation of Computationally Screened Inhibitors. *Journal of Chemical Information and Modeling* **2021**, *61*, 4082–4096.

(36) Hazemann, J.; Kimmerlin, T.; Lange, R.; Sweeney, A. M.; Bourquin, G.; Ritz, D.; Czodrowski, P. Identification of SARS-CoV-2 Mpro inhibitors through deep reinforce-

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

ment learning for de novo drug design and computational chemistry approaches. *bioRxiv* **2024**,

(37) Chenthamarakshan, V. et al. Accelerating drug target inhibitor discovery with a deep generative foundation model. *Science Advances* **2023**, *9*, eadg7865.

(38) Landrum, G. RDKit: Open-source cheminformatics. `http://www.rdkit.org/`.

(39) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574.

(40) Maier, J. A.; Martinez, C.; Kasavajhala, K.; Wickstrom, L.; Hauser, K. E.; Simmerling, C. ff14SB: Improving the Accuracy of Protein Side Chain and Backbone Parameters from ff99SB. *J. Chem. Theory Comput.* **2015**, *11*, 3696–3713.

(41) Boothroyd, S. et al. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. *Journal of Chemical Theory and Computation* **2023**, *19*, 3251–3275.

(42) Wang, J.; Wolf, R. M.; Caldwell, J. W.; Kollman, P. A.; Case, D. A. Development and testing of a general amber force field. *J. Comput. Chem.* **2004**, *25*, 1157–1174.

(43) Ragoza, M.; Hochuli, J.; Idrobo, E.; Sunseri, J.; Koes, D. R. Protein–Ligand Scoring with Convolutional Neural Networks. *J. Chem. Inf. Model.* **2017**, *57*, 942–957.

(44) Francoeur, P. G.; Masuda, T.; Sunseri, J.; Jia, A.; Iovanisci, R. B.; Snyder, I.; Koes, D. R. Three-Dimensional Convolutional Neural Networks and a Cross-Docked Data Set for Structure-Based Drug Design. *J. Chem. Inf. Model.* **2020**, *60*, 4200–4215.

(45) Deringer, V. L.; Bartók, A. P.; Bernstein, N.; Wilkins, D. M.; Ceriotti, M.; Csányi, G. Gaussian Process Regression for Materials and Molecules. *Chemical Reviews* **2021**, *121*, 10073–10141.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

(46) Wang, J. An Intuitive Tutorial to Gaussian Process Regression. *Computing in Science & Engineering* **2023**, *25*, 4–11.

(47) Wang, A.; Liang, H.; McDannald, A.; Takeuchi, I.; Kusne, A. G. Benchmarking active learning strategies for materials optimization and discovery. *Oxford Open Materials Science* **2022**, *2*, itac006.

(48) Ertl, P.; Schuffenhauer, A. Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions. *J. Cheminformatics* **2009**, *1*, 8.

(49) Alnammi, M.; Liu, S.; Ericksen, S. S.; Ananiev, G. E.; Voter, A. F.; Guo, S.; Keck, J. L.; Hoffmann, F. M.; Wildman, S. A.; Gitter, A. Evaluating Scalable Supervised Learning for Synthesize-on-Demand Chemical Libraries. *Journal of Chemical Information and Modeling* **2023**, *63*, 5513–5528.

(50) Pettersen, E. F.; Goddard, T. D.; Huang, C. C.; Couch, G. S.; Greenblatt, D. M.; Meng, E. C.; Ferrin, T. E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.

(51) Takeuchi, K.; Kunimoto, R.; Bajorath, J. R-group replacement database for medicinal chemistry. *Future Sci. OA* **2021**, *7*, 8.

(52) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(53) Danka, T.; Horvath, P. modAL: A modular active learning framework for Python. available on arXiv at `https://arxiv.org/abs/1805.00979`.

(54) Dask Development Team Dask: Library for dynamic task scheduling. 2024.

(55) Noske1, G. D.; de Souza Silva1, E.; de Godoy1, M. O.; Dolci1, I.; Fernandes1, R. S.; Guido1, R. V. C.; Sjö, P.; Oliva1, G.; Godoy, A. S. Structural basis of nirmatrelvir

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0

and ensitrelvir activity against naturally occurring polymorphisms of the SARS-CoV-2 main protease. *Journal of Biological Chemistry* **2023**, *299*, 103004.

(56) Anstine, D.; Zubatyuk, R.; Isayev, O. AIMNet2: A Neural Network Potential to Meet your Neutral, Charged, Organic, and Elemental-Organic Needs. *ChemRxiv* **2024**,

(57) Kovács, D. P.; Moore, J. H.; Browning, N. J.; Batatia, I.; Horton, J. T.; Kapil, V.; Witt, W. C.; Magdau, I.-B.; Cole, D. J.; Csányi, G. MACE-OFF23: Transferable Machine Learning Force Fields for Organic Molecules. 2023.

(58) Fromer, J. C.; Graff, D. E.; Coley, C. W. Pareto optimization to accelerate multi-objective virtual screening. *Digital Discovery* **2024**, *3*, 467–481.

(59) Powers, A. S.; Yu, H. H.; Suriana, P.; Koodli, R. V.; Lu, T.; Paggi, J. M.; Dror, R. O. Geometric Deep Learning for Structure-Based Ligand Design. *ACS Central Science* **2023**, *9*, 2257–2267.

https://doi.org/10.26434/chemrxiv-2024-xczfb **ORCID:** https://orcid.org/0000-0003-2933-0719 Content not peer-reviewed by ChemRxiv. **License:** CC BY 4.0