

Chapter 4

Discovery of SARS-CoV-2 main protease inhibitors via synthesis-directed de novo design

This chapter is based on Aaron Morris, William McCorkindale, The COVID Moonshot Consortium, Nir Drayman, John D. Chodera, Savaş Tay, Nir London and Alpha A. Lee. Discovery of SARS-CoV-2 main protease inhibitors using a synthesis-directed de novo design model, *Chem. Commun.*, 2021,57, 5909-5912

Coronaviruses are a family of pathogens that is frequently associated with serious and highly infectious human diseases, from the common cold to the SARS-CoV pandemic (2003, 774 deaths, 11% fatality rate), MERS-CoV pandemic (2012, 858 deaths, 34% fatality rate) and most recently the COVID-19 pandemic (ongoing pandemic, 1.7 million deaths up to Dec 2020).

The main protease (Mpro) is one of the best characterized drug targets for direct-acting antivirals [139, 25]. Mpro is essential for viral replication and its binding site is distinct from known human proteases, thus inhibitors are unlikely to be toxic [81, 104]. Moreover, the high degree of conservation across different coronaviruses renders Mpro targeting a fruitful avenue towards pan-cornavirus antivirals [186]. To date, most reported Mpro inhibitors are peptidomimetics, covalent, or both [25]. Peptidomimetics are challenging to develop into oral therapeutics, and covalent inhibitors incur additional idiosyncratic toxicity risks. We launched the COVID Moonshot consortium in March 2020, aiming to find oral antivirals against COVID-19 in an open-science, patent-free manner [27].

38 Discovery of SARS-CoV-2 main protease inhibitors via synthesis-directed de novo design

Here we report the prospective use of algorithmic *de novo* design to rapidly expand hits utilising machine learning (ML) models for ranking compounds by bioactivity as well as synthesis route prediction. Starting from 42 compounds with IC₅₀ within assay dynamic range ($< 100\mu\text{M}$) and 515 inactives, our model designed 5 new compounds predicted to have higher activity, together with predicted synthetic routes. All designs were chemically synthesized and experimentally tested, and 3 have measurable activity against Mpro. The top compound has comparable Mpro inhibition to the best in the training set, but with a different scaffold, and is active against the OC43 coronavirus in a live virus assay.

4.1 Learning to rank compounds

Our compound prioritisation model aims to predict whether a designed compound is likely to be an improvement in activity over the incumbent. However, as is typical in the hit-expansion stage, bioactivity modelling is hindered by insufficient data where the majority of compounds are inactive, and noisy data as measurement variability increases for lower affinity compounds. Thresholding the data and framing the problem as classification of active/inactive would not allow us to rank compounds based on predicted improvement over the incumbent, yet the relatively small number of quantitative potency measurement bioactivity data and the measurement noise makes a regression approach challenging.

Instead of predicting IC₅₀ values directly, we focused our attention on a learn-to-rank approach [44, 3] that predicts the pairwise comparison of ligands - given a pair of molecules (A, B), the model predicts whether A is more active than B . This approach allows us to assimilate both coarse (active/inactive) and fine (quantitative potency measurements) data into a single model, effectively combining ‘easy’ classification and ‘difficult’ regression into a ‘moderate’ task of ranking input pairs. With a trained ranking model, we can screen for more potent inhibitors by ranking new molecules against the most potent active compounds in the dataset.

To implement this ranking model (Figure 4.1), we use the difference in molecular fingerprints between two molecules, $f_A - f_B$ as input to the model, and the output is whether the molecule A is more or less potent than molecule B . The descriptor we use for representing a molecule is a concatenation of 3 512-bit fingerprint representations (Morgan, Atom, Topological Torsion) into one 1536 representation. A multilayer perceptron, implemented via FastAI Tabular framework [67] with default hyperparameters, is used for modelling the data. The choice of molecular descriptor and model was based on empirical performance (Details on the model implementation can be found in Appendix A.2.2).

To construct a suitable dataset for training the model, the activity data must be reformatted into pairs of molecules. This is done by pairing all actives with all inactives, as well as pairing

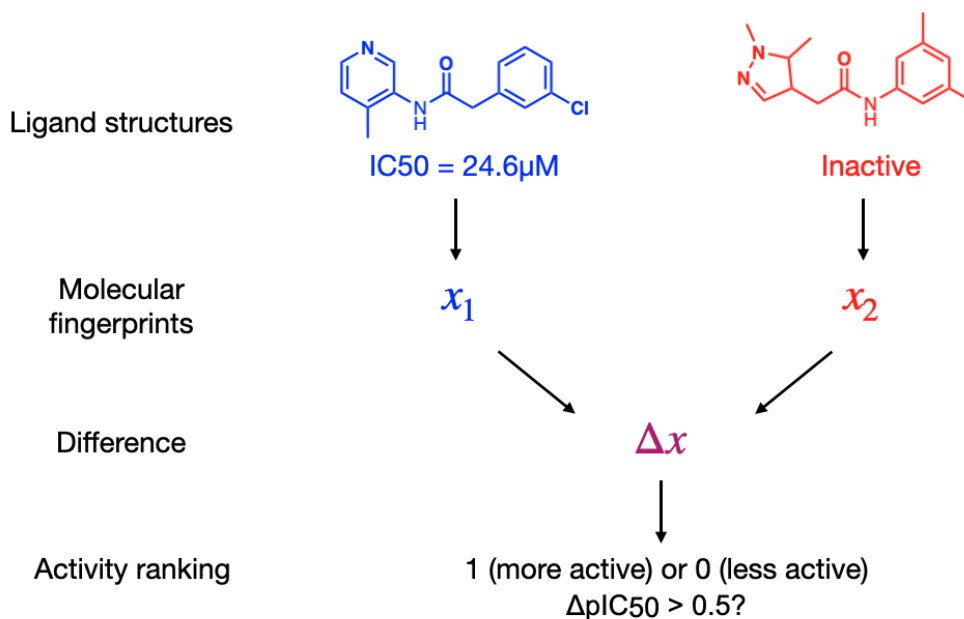


Fig. 4.1 A schematic of the model setup. A classifier takes the difference in molecular fingerprint between two molecules and predicts where one molecule is more or less active than the other.

up actives where there is a significant different in bioactivity ($\Delta pIC_{50} > 0.5$). This threshold was chosen to match typical assay error, forcing the model to ignore irrelevant experimental noise by ensuring that it is ranking only ligands with demonstrably different bioactivity. The inactive molecules were not paired up between each other as the ranking of inactivity is not relevant to the task at hand and potentially noisy/misleading.

A flipped pairing is included for all pairs so that the dataset is antisymmetric since the model would just predict ones otherwise. Theoretically this method is appealing because it is a natural way of oversampling the low proportion of actives, addressing the problem of dataset imbalance commonly seen in drug discovery classification tasks. Additionally, creating pairs between the actives allows the exploitation of activity information without the noise/difficulty of trying to learn accurate pIC_{50} values.

For performance evaluation, the dataset was randomly split into training (80%) and testing (20%) sets (with roughly the same active/inactive proportion) before the molecules are paired up independently within each set. This ensures that there is no cross-talk between the train/test sets where the model could simply memorize the activity of certain compounds. We train the model on pairs of compounds within the training set and evaluate the model on both pairs within the test set as well as pairs between the training and test set.

Figure 4.2 shows that our binary ranking model achieves an AUC of 0.88 (95% CI: [0.83,0.96]) in ranking ligands within the test set, and AUC for 0.94 (95% CI: [0.91,0.98])

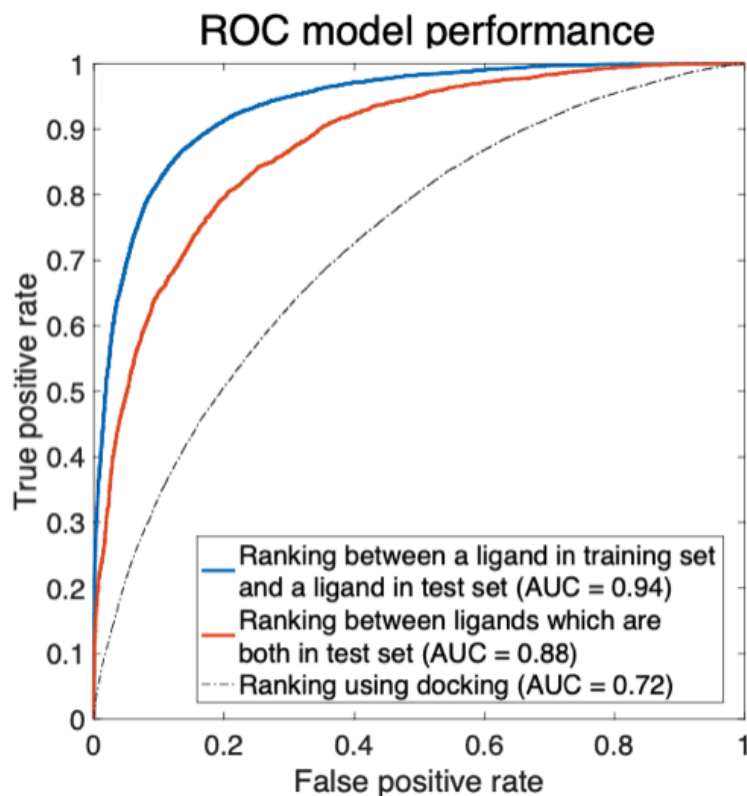


Fig. 4.2 Relative ranking of ligands can be predicted by our learning-to-rank machine learning model. The Receiver Operating Characteristic curve of classifying whether a molecule is more/less active than the other.

where we compare a ligand in the training set against another ligand in the test set; the latter is more relevant as our goal is finding ligands more active than the best incumbent. The 95% confidence interval is computed using bootstrapping. We also compare our model against ranking compounds using docking scores generated with OpenEye's FRED docking algorithm, which achieves an AUC of 0.72 (95% CI: [0.722,0.723]) (Details of the docking procedure can be found in Appendix A.1). Note that docking does not require ligand bioactivity as training data, thus is not a directly comparison to machine learning.

Beyond train-test split, model performance can be evaluated from a time-split. Five months have elapsed from the time we deployed our model for prospective compound selection to writing up the this work. During that time, the COVID Moonshot Consortium (a team of expert medicinal chemists) has independently designed, synthesised and tested 356 compounds [182], out of which 15% were better than the top 2 compounds (having IC_{50} comparable within error) in our dataset. Table 4.1 shows that our model has an enrichment factor of ~ 2 , i.e. if we rescore the 356 compounds synthesized by the medicinal chemistry team using our model, and pick

the top 1%-10% percentile, the proportion of molecules that would be better than the top 2 compounds would be $\sim 2\times$ higher than human selection.

Percentile	1%	2.5%	10%
Enrichment Factor	1.7	2.3	1.7

Table 4.1 Enrichment factor for the time-split dataset, where we consider model performance on data arriving after the model has been deployed to generate compounds for synthesis and testing.

These retrospective results illustrate that a learning-to-rank approach can leverage bioactivity data from both active and inactive molecules for the enrichment of potent compounds in a real-world drug discovery campaign. In the next section, we deploy our model to discover new Mpro inhibitors in a prospective experiment.

4.2 Prospective chemical space exploration

After designing a well-founded ML scoring model, we must decide on a virtual library of compounds to explore. While one could screen an ultra-large library of make-on-demand compounds as in the previous chapter, it is only feasible for a relatively cheap computational model which is not the case for the ML model developed in this work.

Instead, we consider a more targeted approach by exploring the chemical space of chemical substructures contained within our initial dataset. Building on rule-based fragmentation methods such as BRICS [40] and CReM [140], our general approach is to decompose the existing molecules into a large set of distinct substructure components before enumerating all components with one another, generating a large number of novel and diverse molecules.

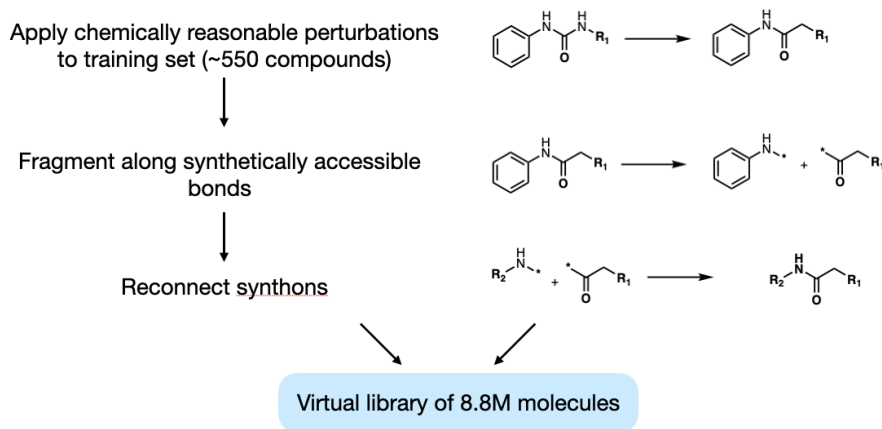


Fig. 4.3 A schematic of the methodology for the library generation process.

42 Discovery of SARS-CoV-2 main protease inhibitors via synthesis-directed de novo design

Specifically, we first introduce a set of chemically reasonable perturbations (linker and chemotype swaps, e.g. amide to retroamide, amide to urea, swapping N-aryl groups), which is applied to the whole set of active molecules. We then fragment along synthetically accessible bonds (e.g. amides and aromatic C-C and C-N), and reconnect the synthons to generate an exhaustive library (Figure 4.3). These operations are defined using SMARTS rules.

The resulting library of 8.8 million generated molecules is then scored against the top 3 compounds in the training set using the learning-to-rank framework, and the mean score is taken as the final score for each compound.

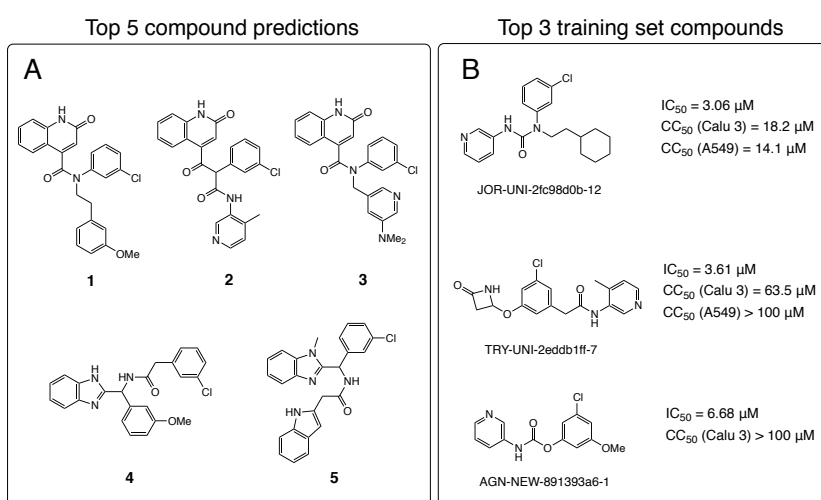


Fig. 4.4 Our synthesis-driven design model prioritises molecular scaffold that are not in the top hits. (A) The 5 compounds selected by our methodology for synthesis and testing. (B) The top 3 compounds from the training set, with potency and cytotoxicity measurements.

Although virtual “reactions” were used to generate new molecules, the synthons are not necessarily off-the-shelf nor the reactions optimal. As such, we use a retrosynthesis predictor to triage based on synthetic accessibility. We used Manifold, a platform for synthesis route prediction (<https://postera.ai/manifold>), to generate synthetic routes for the model’s top-ranked molecules starting from purchasable building blocks. The underlying technology is based on Molecular Transformer, a machine learning model for reaction prediction using sequence-to-sequence translation [100, 165]. The top 5 molecules from the screening library with <4 steps in their predicted routes were synthesised and tested (Figure 4.4A). For comparison, the most potent molecules from the training set are shown in Figure 4.4B. All five compounds have Tanimoto similarity <0.48 (1024-bit ECFP6) to any molecule in the training set, indicating that the model is not merely reproducing molecules similar to the most potent actives but are exploring novel scaffolds.

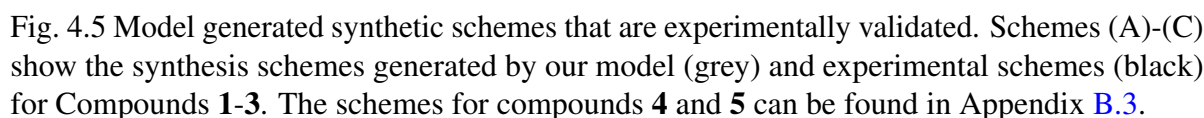


Figure 4.5 shows that for Compounds **1**, **2**, **4** and **5** our retrosynthesis algorithm generates successful routes, thus provides a reasonable estimate of synthetic complexity. The syntheses were carried out at the Wuxi AppTec and compounds were assayed as received. Minor variations in building blocks were employed depending on what was readily available. We note that our algorithm failed to estimate the synthetic complexity of Compound **3**. The final amide formation step was unexpectedly challenging, and no desired product was seen despite significant efforts in condition screening. Compound **3** was furnished via an alternative strategy, employing an Ullmann coupling to arylate the amide, which was not predicted by our approach.

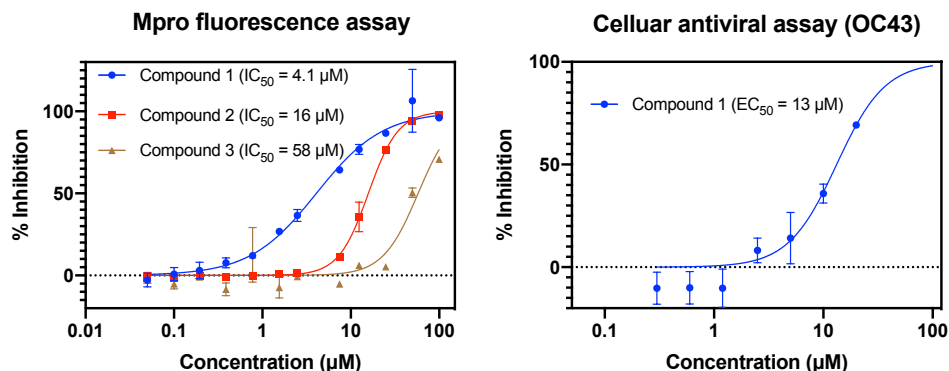
44 Discovery of SARS-CoV-2 main protease inhibitors via synthesis-directed de novo design

Fig. 4.6 Three compounds generated using our synthesis-directed model exhibit Mpro activity. Our most active compound has measurable antiviral activity against the OC43 coronavirus and no measurable cytotoxic effect.

Compounds **1-5** were tested for Mpro activity using a fluorescence assay. Figure 4.6 shows that Compounds **1-3** have IC₅₀ within assay dynamic range ($< 100\mu\text{M}$), and Compound **1** has IC₅₀ = $4.1\mu\text{M}$ (95% CI: [3.42,4.86]). Compound **1** is further assayed in live virus assays, with the less pathogenic OC43 coronavirus, showing EC₅₀ = $13\mu\text{M}$ (95% CI: [10.1, 18.4]) and is not cytotoxic (CC₅₀ $> 100\mu\text{M}$ against A549 cell line; CC₅₀ is the concentration required to cause 50% cell death). We employ OC43 as a rapid surrogate assay for SARS-CoV-2 as the former can be done in a BSL-2 rather than BSL-3 lab (See Appendix B for assay details).

4.3 Discussion

In summary, we demonstrated the utility of a *de novo* design framework that learns to rank bioactivity and estimates synthetic complexity, for generating ideas in hit expansion. At the time of writing, optimisation of a quinolone-based scaffold is ongoing in the COVID Moonshot initiative (<https://postera.ai/covid>). Data for Compound **1-5** is registered as the ALP-POS-ddb41b15 series on the Moonshot platform.

The learning-to-rank approach presented in this chapter is a promising technique for maximally utilising data from inactive compounds as well as accounting for noise in the experimental values. Further work extending this approach to utilise information from ligand-protein complexes has shown great success in hit-to-lead optimisation [151]. By using docking-based structural descriptors as the input to their ranking model, the authors were able to utilise crystal structures of inactive compounds and outperform docking as well as a fingerprint-based ranking model on ranking ligand activity. Applying this model on a docked structures from a virtual library, the authors were able to greatly improve the potency from their starting point and

extended the ligand into unknown regions of the binding site. This result shows the potential for applying ranking-based approaches for modelling bioactivity.

A caveat for pair-based ranking is that even for moderately sized datasets the number of molecular pairs (which is proportional to the square of the dataset size) may become very large and therefore unfeasible to train on. Engineering approaches for efficient sampling of molecular pairs, such as using Tanimoto similarity to constrain pair selection, may be necessary and should be the subject of future work.

The AI-based estimation of synthetic feasibility played a critical role in compound selection in this work and is a promising enabler for automated design workflows in drug discovery [55]. Beyond the capability to triage libraries by synthesisability and to guide the design of synthetic routes, ML synthesis models could play a key part in an automated drug design workflow. Utilising an automated workflow for the automatic generation of molecular designs as well as decision-making could potentially reduce iteration cycle time, require fewer compounds and iterations to produce a candidate, and scale to more programs [158, 32]. The framework presented in this chapter is an example of an automated design workflow albeit only with a single iteration - utilisation of multiple iterations via an optimisation [95] or reinforcement learning feedback loop [18] will be a challenging but exciting area of future work.

References

- [Che] Chemspace: Lead-like compounds. 1
- [SMA] Smarts - a language for describing molecular patterns. 2
- [3] Agarwal, S., Dugar, D., and Sengupta, S. (2010). Ranking chemical structures for drug discovery: a new machine learning approach. *Journal of chemical information and modeling*, 50(5):716–731. 3
- [4] Allen, T. E. H., Wedlake, A. J., Gelžinytė, E., Gong, C., Goodman, J. M., Gutsell, S., and Russell, P. J. (2020). Neural network activation similarity: a new measure to assist decision making in chemical toxicology. *Chem. Sci.*, 11:7335–7348. 4
- [5] Alon, A., Lyu, J., Braz, J. M., Tummino, T. A., Craik, V., O’Meara, M. J., Webb, C. M., Radchenko, D. S., Moroz, Y. S., Huang, X.-P., Liu, Y., Roth, B. L., Irwin, J. J., Basbaum, A. I., Shoichet, B. K., and Kruse, A. C. (2021). Structures of the 5
- σ
- 2 receptor enable docking for bioactive ligand discovery. *Nature*, 600(7890):759–764. 6
- [6] Antunes, D. A., Devaurs, D., and Kavraki, L. E. (2015). Understanding the challenges of protein flexibility in drug design. *Expert Opinion on Drug Discovery*, 10(12):1301–1313. 7
- [7] Baell, J. B. and Holloway, G. A. (2010). New substructure filters for removal of pan assay interference compounds (pains) from screening libraries and for their exclusion in bioassays. *Journal of Medicinal Chemistry*, 53(7):2719–2740. PMID: 20131845. 8
- [8] Bajusz, D., Rácz, A., and Héberger, K. (2015). Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of Cheminformatics*, 7(1):20. 9
- [9] Baldi, P. and Nasr, R. (2010). When is chemical similarity significant? the statistical distribution of chemical similarity scores and its extreme values. *Journal of Chemical Information and Modeling*, 50(7):1205–1222. 10
- [10] Ballante, F., Kooistra, A. J., Kampen, S., de Graaf, C., and Carlsson, J. (2021). Structure-based virtual screening for ligands of g protein-coupled receptors: What can molecular docking do for you? *Pharmacological Reviews*, 73(4):1698–1736. 11
- [11] Baydin, A. G., Pearlmutter, B. A., Radul, A. A., and Siskind, J. M. (2018). Automatic differentiation in machine learning: a survey. *Journal of Machine Learning Research*, 18(153):1–43. 12

- [12] Bender, A. and Cortés-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 1: Ways to make an impact, and why we are not there yet. *Drug Discovery Today*, 26(2):511–524.
- [13] Bender, A. and Cortes-Ciriano, I. (2021). Artificial intelligence in drug discovery: what is realistic, what are illusions? part 2: a discussion of chemical and biological data. *Drug Discovery Today*, 26(4):1040–1052.
- [14] Bender, B. J., Gahbauer, S., Lutten, A., Lyu, J., Webb, C. M., Stein, R. M., Fink, E. A., Balius, T. E., Carlsson, J., Irwin, J. J., and Shoichet, B. K. (2021). A practical guide to large-scale docking. *Nature Protocols*, 16(10):4799–4832.
- [15] Bishop, C. M. (2006). *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.
- [16] Bjerrum, E. J. (2017). Smiles enumeration as data augmentation for neural network modeling of molecules.
- [17] Blakemore, D. C., Castro, L., Churcher, I., Rees, D. C., Thomas, A. W., Wilson, D. M., and Wood, A. (2018). Organic synthesis provides opportunities to transform drug discovery. *Nature chemistry*, 10(4):383.
- [18] Born, J., Manica, M., Oskooei, A., Cadow, J., Borgwardt, K., and Martínez, M. R. (2019). Paccmann^{RL}: Designing anticancer drugs from transcriptomic data via reinforcement learning.
- [19] Boström, J., Brown, D. G., Young, R. J., and Keserü, G. M. (2018). Expanding the medicinal chemistry synthetic toolbox. *Nature Reviews Drug Discovery*.
- [20] Botev, Z. I., Grotowski, J. F., and Kroese, D. P. (2010). Kernel density estimation via diffusion. *The Annals of Statistics*, 38(5):2916 – 2957.
- [21] Bradshaw, J., Kusner, M. J., Paige, B., Segler, M. H. S., and Hernández-Lobato, J. M. (2019). A generative model for electron paths.
- [22] Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.
- [23] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners.
- [24] Butina, D. (1999). Unsupervised data base clustering based on daylight’s fingerprint and tanimoto similarity: A fast and automated way to cluster small and large data sets. *Journal of Chemical Information and Computer Sciences*, 39(4):747–750.
- [25] Cannalire, R., Cerchia, C., Beccari, A. R., Di Leva, F. S., and Summa, V. (2020). Targeting sars-cov-2 proteases and polymerase for covid-19 treatment: State of the art and future opportunities. *Journal of medicinal chemistry*.

- [26] Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.
- [27] Chodera, J., Lee, A. A., London, N., and von Delft, F. (2020). Crowdsourcing drug discovery for pandemics. *Nature Chemistry*, 12(7):581–581.
- [28] Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling.
- [29] Clayden, J., Greeves, N., and Warren, S. (2012). *Organic Chemistry*. Oxford University Press, 2nd edition.
- [30] Coleman, R. G., Carchia, M., Sterling, T., Irwin, J. J., and Shoichet, B. K. (2013). Ligand pose and orientational sampling in molecular docking. *PLOS ONE*, 8(10):1–19.
- [31] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2019a). Autonomous Discovery in the Chemical Sciences Part I: Progress. *Angewandte Chemie - International Edition*, pages 2–38.
- [32] Coley, C. W., Eyke, N. S., and Jensen, K. F. (2020). Autonomous discovery in the chemical sciences part ii: Outlook. *Angewandte Chemie International Edition*, 59(52):23414–23436.
- [33] Coley, C. W., Green, W. H., and Jensen, K. F. (2018). Machine Learning in Computer-Aided Synthesis Planning. *Accounts of Chemical Research*, 51(5):1281–1289.
- [34] Coley, C. W., Jin, W., Rogers, L., Jamison, T. F., Jaakkola, T. S., Green, W. H., Barzilay, R., and Jensen, K. F. (2019b). A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.*, 10:370–377.
- [35] Commons, W. (2018). Receiver operating characteristic (roc) curve. https://commons.wikimedia.org/wiki/File:Roc_curve.svg.
- [36] Corey, E. J., Long, A. K., and Rubenstein, S. D. (1985). Computer-assisted analysis in organic synthesis. *Science*, 228(4698):408.
- [37] Corso, G., Stärk, H., Jing, B., Barzilay, R., and Jaakkola, T. S. (2023). Diffdock: Diffusion steps, twists, and turns for molecular docking. In *The Eleventh International Conference on Learning Representations*.
- [38] Cortés-Ciriano, I. and Bender, A. (2019). Reliable prediction errors for deep neural networks using test-time dropout. *Journal of Chemical Information and Modeling*, 59(7):3330–3339. PMID: 31241929.
- [39] Davis, B. J. and Roughley, S. D. (2017). Chapter eleven - fragment-based lead discovery. In Goodnow, R. A., editor, *Platform Technologies in Drug Discovery and Validation*, volume 50 of *Annual Reports in Medicinal Chemistry*, pages 371–439. Academic Press.
- [40] Degen, J., Wegscheid-Gerlach, C., Zaliani, A., and Rarey, M. (2008). On the art of compiling and using ‘drug-like’ chemical fragment spaces. *ChemMedChem*, 3(10):1503–1507.

- [41] Delaney, J. S. (2004). Esol: Estimating aqueous solubility directly from molecular structure. *Journal of Chemical Information and Computer Sciences*, 44(3):1000–1005.
- [42] Dixon, S. L., Smondryev, A. M., and Rao, S. N. (2006). Phase: A novel approach to pharmacophore modeling and 3d database searching. *Chemical Biology & Drug Design*, 67(5):370–372.
- [43] Douangamath, A., Fearon, D., Gehrtz, P., Krojer, T., Lukacik, P., Owen, C. D., Resnick, E., Strain-Damerell, C., Aimon, A., Ábrányi-Balogh, P., Brandão-Neto, J., Carbery, A., Davison, G., Dias, A., Downes, T. D., Dunnett, L., Fairhead, M., Firth, J. D., Jones, S. P., Keeley, A., Keserü, G. M., Klein, H. F., Martin, M. P., Noble, M. E. M., O'Brien, P., Powell, A., Reddi, R. N., Skyner, R., Snee, M., Waring, M. J., Wild, C., London, N., von Delft, F., and Walsh, M. A. (2020). Crystallographic and electrophilic fragment screening of the sars-cov-2 main protease. *Nature Communications*, 11(1):5047.
- [44] Duffy, N. P. (2010). Molecular property modeling using ranking. US Patent 7,702,467.
- [45] Eberhardt, J., Santos-Martins, D., Tillack, A. F., and Forli, S. (2021). Autodock vina 1.2.0: New docking methods, expanded force field, and python bindings. *Journal of Chemical Information and Modeling*, 61(8):3891–3898.
- [46] Erickson, J. A., Jalaie, M., Robertson, D. H., Lewis, R. A., and Vieth, M. (2004). Lessons in molecular recognition: The effects of ligand and protein flexibility on molecular docking accuracy. *Journal of Medicinal Chemistry*, 47(1):45–55.
- [47] Fink, E. A., Xu, J., Hübner, H., Braz, J. M., Seemann, P., Avet, C., Craik, V., Weikert, D., Schmidt, M. F., Webb, C. M., Tolmachova, N. A., Moroz, Y. S., Huang, X.-P., Kalyanaraman, C., Gahbauer, S., Chen, G., Liu, Z., Jacobson, M. P., Irwin, J. J., Bouvier, M., Du, Y., Shoichet, B. K., Basbaum, A. I., and Gmeiner, P. (2022). Structure-based discovery of nonopioid analgesics acting through the α_{2A} -adrenergic receptor. *Science*, 377(6614):eabn7065.
- [48] Friedel, C. and Crafts, J. (1877). Sur une nouvelle méthode générale de synthèse d'hydrocarbures, d'acétones, etc.
- [49] Friesner, R. A., Banks, J. L., Murphy, R. B., Halgren, T. A., Klicic, J. J., Mainz, D. T., Repasky, M. P., Knoll, E. H., Shelley, M., Perry, J. K., Shaw, D. E., Francis, P., and Shenkin, P. S. (2004). Glide: A new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of Medicinal Chemistry*, 47(7):1739–1749.
- [50] Gehrtz, P., Marom, S., Bührmann, M., Hardick, J., Kleinbölting, S., Shraga, A., Dubiella, C., Gabizon, R., Wiese, J. N., Müller, M. P., Cohen, G., Babaev, I., Shurrush, K., Avram, L., Resnick, E., Barr, H., Rauh, D., and London, N. (2022). Optimization of covalent mkk7 inhibitors via crude nanomole-scale libraries. *Journal of Medicinal Chemistry*, 65(15):10341–10356.
- [51] Gentile, F., Agrawal, V., Hsing, M., Ton, A.-T., Ban, F., Norinder, U., Gleave, M. E., and Cherkasov, A. (2020). Deep docking: A deep learning platform for augmentation of structure based drug discovery. *ACS Central Science*, 6(6):939–949.

- [52] Gesmundo, N. J., Sauvagnat, B., Curran, P. J., Richards, M. P., Andrews, C. L., Dandliker, P. J., and Cernak, T. (2018). Nanoscale synthesis and affinity ranking. *Nature*, 557(7704):228–232.
- [53] Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. (2017). Neural message passing for quantum chemistry. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML’17, page 1263–1272. JMLR.org.
- [54] Girona-Martínez, A., Donckele, E. J., Samain, F., and Neri, D. (2021). Dna-encoded chemical libraries: A comprehensive review with succesful stories and future challenges. *ACS Pharmacology & Translational Science*, 4(4):1265–1279.
- [55] Goldman, B., Kearnes, S., Kramer, T., Riley, P., and Walters, W. P. (2022). Defining levels of automated chemical design. *Journal of Medicinal Chemistry*, 65(10):7073–7087.
- [56] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press. <http://www.deeplearningbook.org>.
- [57] Gorgulla, C., Boeszoermenyi, A., Wang, Z.-F., Fischer, P. D., Coote, P. W., Padmanabha Das, K. M., Malets, Y. S., Radchenko, D. S., Moroz, Y. S., Scott, D. A., Fackeldey, K., Hoffmann, M., Iavniuk, I., Wagner, G., and Arthanari, H. (2020). An open-source drug discovery platform enables ultra-large virtual screens. *Nature*, 580(7805):663–668.
- [58] Griffiths, R.-R., Greenfield, J. L., Thawani, A. R., Jamasb, A. R., Moss, H. B., Bourached, A., Jones, P., McCorkindale, W., Aldrick, A. A., Fuchter, M. J., and Lee, A. A. (2022). Data-driven discovery of molecular photoswitches with multioutput gaussian processes. *Chem. Sci.*, 13:13541–13551.
- [59] Guan, Y., Coley, C. W., Wu, H., Ranasinghe, D., Heid, E., Struble, T. J., Pattanaik, L., Green, W. H., and Jensen, K. F. (2021). Regio-selectivity prediction with a machine-learned reaction representation and on-the-fly quantum mechanical descriptors. *Chem. Sci.*, 12:2198–2208.
- [60] Hall, R. J., Murray, C. W., and Verdonk, M. L. (2017). The fragment network: A chemistry recommendation engine built using a graph database. *Journal of Medicinal Chemistry*, 60(14):6440–6450.
- [61] Hann, M. M., Leach, A. R., and Harper, G. (2001). Molecular complexity and its impact on the probability of finding leads for drug discovery. *Journal of chemical information and computer sciences*, 41(3):856–864.
- [62] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.
- [63] Heller, S., McNaught, A., Stein, S., Tchekhovskoi, D., and Pletnev, I. (2013). Inchi - the worldwide chemical structure identifier standard. *Journal of Cheminformatics*, 5(1):7.
- [64] Hermann, J. C., Chen, Y., Wartchow, C., Menke, J., Gao, L., Gleason, S. K., Haynes, N.-E., Scott, N., Petersen, A., Gabriel, S., Vu, B., George, K. M., Narayanan, A., Li, S. H., Qian, H., Beatini, N., Niu, L., and Gan, Q.-F. (2013). Metal impurities cause false positives in high-throughput screening campaigns. *ACS Medicinal Chemistry Letters*, 4(2):197–200.

- [65] Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.
- [66] Hofmann, T., Schölkopf, B., and Smola, A. J. (2008). Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171 – 1220.
- [67] Howard, J. et al. (2018). fastai. <https://github.com/fastai/fastai>.
- [68] Huang, N., Shoichet, B. K., and Irwin, J. J. (2006). Benchmarking sets for molecular docking. *Journal of Medicinal Chemistry*, 49(23):6789–6801.
- [69] Hughes, J. P., Rees, S., Kalindjian, S. B., and Philpott, K. L. (2011). Principles of early drug discovery. *British journal of pharmacology*, 162(6):1239–1249.
- [70] Ichihara, O., Barker, J., Law, R. J., and Whittaker, M. (2011). Compound design by fragment-linking. *Molecular Informatics*, 30(4):298–306.
- [71] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2018). Protein family-specific models using deep neural networks and transfer learning improve virtual screening and highlight the need for more data. *Journal of Chemical Information and Modeling*, 58(11):2319–2330.
- [72] Imrie, F., Bradley, A. R., van der Schaar, M., and Deane, C. M. (2020). Deep generative models for 3d linker design. *Journal of Chemical Information and Modeling*, 60(4):1983–1995.
- [73] Imrie, F., Hadfield, T. E., Bradley, A. R., and Deane, C. M. (2021). Deep generative design with 3d pharmacophoric constraints. *Chem. Sci.*, 12:14577–14589.
- [74] Ioffe, S. and Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift.
- [75] Irwin, J. J., Tang, K. G., Young, J., Dandarchuluun, C., Wong, B. R., Khurelbaatar, M., Moroz, Y. S., Mayfield, J., and Sayle, R. A. (2020). Zinc20—a free ultralarge-scale chemical database for ligand discovery. *Journal of Chemical Information and Modeling*, 60(12):6065–6073.
- [76] Jia, X., Lynch, A., Huang, Y., Danielson, M., Lang’at, I., Milder, A., Ruby, A. E., Wang, H., Friedler, S. A., Norquist, A. J., and Schrier, J. (2019). Anthropogenic biases in chemical reaction data hinder exploratory inorganic synthesis. *Nature*, 573:251–255.
- [77] Jiang, D., Wu, Z., Hsieh, C.-Y., Chen, G., Liao, B., Wang, Z., Shen, C., Cao, D., Wu, J., and Hou, T. (2021). Could graph neural networks learn better molecular representation for drug discovery? a comparison study of descriptor-based and graph-based models. *Journal of Cheminformatics*, 13(1):12.
- [78] Jiménez, J., Škalič, M., Martínez-Rosell, G., and De Fabritiis, G. (2018). Kdeep: Protein–ligand absolute binding affinity prediction via 3d-convolutional neural networks. *Journal of Chemical Information and Modeling*, 58(2):287–296.
- [79] Jiménez-Luna, J., Grisoni, F., and Schneider, G. (2020). Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584.

- [80] Jin, W., Coley, C. W., Barzilay, R., and Jaakkola, T. (2017). Predicting organic reaction outcomes with weisfeiler-lehman network. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):2608–2617.
- [81] Jin, Z., Du, X., Xu, Y., Deng, Y., Liu, M., Zhao, Y., Zhang, B., Li, X., Zhang, L., Peng, C., et al. (2020). Structure of mpro from sars-cov-2 and discovery of its inhibitors. *Nature*, 582(7811):289–293.
- [82] Johansson, S., Thakkar, A., Kogej, T., Bjerrum, E., Genheden, S., Bastys, T., Kannas, C., Schliep, A., Chen, H., and Engkvist, O. (2020). Ai-assisted synthesis prediction. *Drug Discovery Today: Technologies*.
- [83] Johnson, M. A. and Maggiora, G. M. (1990). *Concepts and applications of molecular similarity*. Wiley.
- [84] Jorner, K., Brinck, T., Norrby, P.-O., and Buttar, D. (2021). Machine learning meets mechanistic modelling for accurate prediction of experimental activation energies. *Chem. Sci.*, 12:1163–1175.
- [85] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.
- [86] Karpov, P., Godin, G., and Tetko, I. V. (2020). Transformer-CNN: Swiss knife for QSAR modeling and interpretation. *Journal of Cheminformatics*, 12(1):17.
- [87] Kaserer, T., Beck, K. R., Akram, M., Odermatt, A., and Schuster, D. (2015). Pharmacophore models and pharmacophore-based virtual screening: Concepts and applications exemplified on hydroxysteroid dehydrogenases. *Molecules*, 20(12):22799–22832.
- [88] Kearnes, S. (2021). Pursuing a prospective perspective. *Trends in Chemistry*, 3(2):77–79.
- [89] Kim, H., Na, J., and Lee, W. B. (2021). Generative chemical transformer: Neural machine learning of molecular geometric structures from chemical language via attention. *Journal of Chemical Information and Modeling*, 61(12):5804–5814.
- [90] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.
- [91] Kishimoto, A., Buesser, B., Chen, B., and Botea, A. (2019). Depth-first proof-number search with heuristic edge cost and application to chemical synthesis planning. In *Advances in Neural Information Processing Systems*, pages 7224–7234.
- [92] Kitchen, D. B., Decornez, H., Furr, J. R., and Bajorath, J. (2004). Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949.
- [93] Klein, G., Kim, Y., Senellart, J., and Rush, A. M. (2017). OpenNMT.

- [94] Klucznik, T., Mikulak-Klucznik, B., McCormack, M. P., Lima, H., Szymkuć, S., Bhowmick, M., Molga, K., Zhou, Y., Rickershauser, L., Gajewska, E. P., Touthkine, A., Dittwald, P., Startek, M. P., Kirkovits, G. J., Roszak, R., Adamski, A., Sieredzińska, B., Mrksich, M., Trice, S. L., and Grzybowski, B. A. (2018). Efficient Syntheses of Diverse, Medicinally Relevant Targets Planned by Computer and Executed in the Laboratory. *Chem*, 4(3):522–532.
- [95] Korovina, K., Xu, S., Kandasamy, K., Neiswanger, W., Poczos, B., Schneider, J., and Xing, E. P. (2019). Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations.
- [Kovacs et al.] Kovacs, D. P., McCorkindale, W., and Lee, A. A. Molecular Transformer Explainer. <https://github.com/davkovacs/MTEexplainer.git>.
- [97] Krenn, M., Häse, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. (2020). Self-referencing embedded strings (selfies): A 100string representation. *Machine Learning: Science and Technology*, 1(4):045024.
- [98] Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc.
- [99] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. (2019). Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1096.
- [100] Lee, A. A., Yang, Q., Sresht, V., Bolgar, P., Hou, X., Klug-McLeod, J. L., Butler, C. R., et al. (2019). Molecular transformer unifies reaction prediction and retrosynthesis across pharma chemical space. *Chemical Communications*, 55(81):12152–12155.
- [101] Li, J., Fu, A., and Zhang, L. (2019). An overview of scoring functions used for protein–ligand interactions in molecular docking. *Interdisciplinary Sciences: Computational Life Sciences*, 11(2):320–328.
- [102] Limban, C., Nuță, D. C., Chiriță, C., Negreș, S., Arsene, A. L., Goumenou, M., Karakitsios, S. P., Tsatsakis, A. M., and Sarigiannis, D. A. (2018). The use of structural alerts to avoid the toxicity of pharmaceuticals. *Toxicology Reports*, 5:943–953.
- [103] Lipton, Z. C., Berkowitz, J., and Elkan, C. (2015). A critical review of recurrent neural networks for sequence learning.
- [104] Liu, Y., Liang, C., Xin, L., Ren, X., Tian, L., Ju, X., Li, H., Wang, Y., Zhao, Q., Liu, H., et al. (2020). The development of coronavirus 3c-like protease (3clpro) inhibitors from 2010 to 2020. *European journal of medicinal chemistry*, page 112711.
- [105] Liu, Z., Li, Y., Han, L., Li, J., Liu, J., Zhao, Z., Nie, W., Liu, Y., and Wang, R. (2014). PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics*, 31(3):405–412.

- [106] Llanos, M. A., Gantner, M. E., Rodriguez, S., Alberca, L. N., Bellera, C. L., Talevi, A., and Gavernet, L. (2021). Strengths and weaknesses of docking simulations in the sars-cov-2 era: the main protease (mpro) case study. *Journal of Chemical Information and Modeling*, 61(8):3758–3770. PMID: 34313128. 1
2
3
4
- [107] Lluch, A. M., Sánchez-Baeza, F., Messeguer, A., Fusco, C., and Curci, R. (1993). Regio- and chemoselective epoxidation of fluorinated monoterpenes and sesquiterpenes by dioxiranes. *Tetrahedron*, 49(28):6299–6308. 5
6
7
- [108] Lowe, D. M. (2012). *Extraction of chemical structures and reactions from the literature*. Phd, University of Cambridge. 8
9
- [109] Lundberg, S. M. and Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 2017-Decem(Section 2):4766–4775. 10
11
12
- [110] Lyu, J., Irwin, J. J., and Shoichet, B. K. (2023). Modeling the expansion of virtual screening libraries. *Nature Chemical Biology*. 13
14
- [111] Lyu, J., Wang, S., Balias, T. E., Singh, I., Levit, A., Moroz, Y. S., O’Meara, M. J., Che, T., Alga, E., Tolmachova, K., Tolmachev, A. A., Shoichet, B. K., Roth, B. L., and Irwin, J. J. (2019). Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229. 15
16
17
18
- [112] Macip, G., Garcia-Segura, P., Mestres-Truyol, J., Saldivar-Espinoza, B., Ojeda-Montes, M. J., Gimeno, A., Cereto-Massagué, A., Garcia-Vallvé, S., and Pujadas, G. (2022). Haste makes waste: A critical review of docking-based virtual screening in drug repurposing for sars-cov-2 main protease (m-pro) inhibition. *Medicinal research reviews*, 42(2):744–769. 19
20
21
22
- [113] Maggiora, G., Vogt, M., Stumpfe, D., and Bajorath, J. (2014). Molecular similarity in medicinal chemistry. *Journal of Medicinal Chemistry*, 57(8):3186–3204. 23
24
- [114] Maggiora, G. M. (2006). On outliers and activity cliffs why qsar often disappoints. *Journal of Chemical Information and Modeling*, 46(4):1535–1535. 25
26
- [115] Martin, Y. C., Kofron, J. L., and Traphagen, L. M. (2002). Do structurally similar molecules have similar biological activity? *Journal of Medicinal Chemistry*, 45(19):4350–4358. 27
28
29
- [116] Mayr, A., Klambauer, G., Unterthiner, T., Steijaert, M., Wegner, J. K., Ceulemans, H., Clevert, D.-A., and Hochreiter, S. (2018). Large-scale comparison of machine learning methods for drug target prediction on chembl. *Chem. Sci.*, 9:5441–5451. 30
31
32
- [117] McCloskey, K., Sigel, E. A., Kearnes, S., Xue, L., Tian, X., Moccia, D., Gikunju, D., Bazzaz, S., Chan, B., Clark, M. A., Cuzzo, J. W., Guie, M.-A., Guilinger, J. P., Huguet, C., Hupp, C. D., Keefe, A. D., Mulhern, C. J., Zhang, Y., and Riley, P. (2020). Machine learning on dna-encoded libraries: A new paradigm for hit finding. *Journal of Medicinal Chemistry*, 63(16):8857–8866. 33
34
35
36
37
- [118] McCloskey, K., Taly, A., Monti, F., Brenner, M. P., and Colwell, L. J. (2019). Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences of the United States of America*, 116(24):11624–11629. 38
39
40

- [119] McCorkindale, W., Poelking, C., and Lee, A. A. (2020). Investigating 3d atomic environments for enhanced qsar.
- [120] McGann, M. (2012). Fred and hybrid docking performance on standardized datasets. *Journal of Computer-Aided Molecular Design*, 26(8):897–906.
- [121] Meng, X.-Y., Zhang, H.-X., Mezei, M., and Cui, M. (2011). Molecular docking: a powerful approach for structure-based drug discovery. *Current computer-aided drug design*, 7(2):146–157.
- [122] Merget, B., Turk, S., Eid, S., Rippmann, F., and Fulle, S. (2017). Profiling prediction of kinase inhibitors: Toward the virtual assay. *Journal of Medicinal Chemistry*, 60(1):474–485.
- [123] Montavon, G., Samek, W., and Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing: A Review Journal*, 73:1–15.
- [124] Morgan, H. L. (1965). The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113.
- [125] Morreale, F. E., Testa, A., Chaugule, V. K., Bortoluzzi, A., Ciulli, A., and Walden, H. (2017). Mind the metal: A fragment library-derived zinc impurity binds the e2 ubiquitin-conjugating enzyme ube2t and induces structural rearrangements. *Journal of Medicinal Chemistry*, 60(19):8183–8191.
- [126] Morris, A., McCorkindale, W., Consortium, T. C. M., Drayman, N., Chodera, J. D., Tay, S., London, N., and Lee, A. A. (2021). Discovery of sars-cov-2 main protease inhibitors using a synthesis-directed de novo design model. *Chem. Commun.*, 57:5909–5912.
- [127] Mudrakarta, P. K., Taly, A., Sundararajan, M., and Dhamdhare, K. (2018). Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, Melbourne, Australia. Association for Computational Linguistics.
- [128] Muratov, E. N., Bajorath, J., Sheridan, R. P., Tetko, I. V., Filimonov, D., Poroikov, V., Oprea, T. I., Baskin, I. I., Varnek, A., Roitberg, A., et al. (2020). Qsar without borders. *Chemical Society Reviews*, 49(11):3525–3564.
- [129] Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (dud-e): Better ligands and decoys for better benchmarking. *Journal of Medicinal Chemistry*, 55(14):6582–6594.
- [130] Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.
- [131] O’Boyle, N. M. and Sayle, R. A. (2016). Comparing structural fingerprints using a literature-based similarity benchmark. *Journal of Cheminformatics*, 8(1):36.
- [132] Obrezanova, O., Csányi, G., Gola, J. M. R., and Segall, M. D. (2007). Gaussian processes: A method for automatic qsar modeling of adme properties. *Journal of Chemical Information and Modeling*, 47(5):1847–1857.

- [133] Owen, D. R., Allerton, C. M. N., Anderson, A. S., Aschenbrenner, L., Avery, M., Berritt, S., Boras, B., Cardin, R. D., Carlo, A., Coffman, K. J., Dantonio, A., Di, L., Eng, H., Ferre, R., Gajiwala, K. S., Gibson, S. A., Greasley, S. E., Hurst, B. L., Kadar, E. P., Kalgutkar, A. S., Lee, J. C., Lee, J., Liu, W., Mason, S. W., Noell, S., Novak, J. J., Obach, R. S., Ogilvie, K., Patel, N. C., Pettersson, M., Rai, D. K., Reese, M. R., Sammons, M. F., Sathish, J. G., Singh, R. S. P., Steppan, C. M., Stewart, A. E., Tuttle, J. B., Updyke, L., Verhoest, P. R., Wei, L., Yang, Q., and Zhu, Y. (2021). An oral sars-cov-2 m^{pro} inhibitor clinical candidate for the treatment of covid-19. *Science*, 374(6575):1586–1593.
- [134] Pal, S., Kumar, V., Kundu, B., Bhattacharya, D., Preethy, N., Reddy, M. P., and Talukdar, A. (2019). Ligand-based pharmacophore modeling, virtual screening and molecular docking studies for discovery of potential topoisomerase i inhibitors. *Computational and Structural Biotechnology Journal*, 17:291–310.
- [135] Palmer, D. S., O’Boyle, N. M., Glen, R. C., and Mitchell, J. B. O. (2007). Random forest models to predict aqueous solubility. *Journal of Chemical Information and Modeling*, 47(1):150–158.
- [136] Parks, C. D., Gaieb, Z., Chiu, M., Yang, H., Shao, C., Walters, W. P., Jansen, J. M., McGaughey, G., Lewis, R. A., Bembenek, S. D., Ameriks, M. K., Mirzadegan, T., Burley, S. K., Amaro, R. E., and Gilson, M. K. (2020). D3r grand challenge 4: blind prediction of protein–ligand poses, affinity rankings, and relative binding free energies. *Journal of Computer-Aided Molecular Design*, 34(2):99–119.
- [137] Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, 33(3):1065 – 1076.
- [138] Perera, D., Tucker, J. W., Brahmabhatt, S., Helal, C. J., Chong, A., Farrell, W., Richardson, P., and Sach, N. W. (2018). A platform for automated nanomole-scale reaction screening and micromole-scale synthesis in flow. *Science*, 359(6374):429–434.
- [139] Pillaiyar, T., Manickam, M., Namasivayam, V., Hayashi, Y., and Jung, S.-H. (2016). An overview of severe acute respiratory syndrome–coronavirus (sars-cov) 3cl protease inhibitors: peptidomimetics and small molecule chemotherapy. *Journal of medicinal chemistry*, 59(14):6595–6628.
- [140] Polishchuk, P. (2020). Crem: chemically reasonable mutations framework for structure generation. *Journal of Cheminformatics*, 12(1):28.
- [141] Polishchuk, P. G., Muratov, E. N., Artemenko, A. G., Kolumbin, O. G., Muratov, N. N., and Kuz’min, V. E. (2009). Application of random forest approach to qsar prediction of aquatic toxicity. *Journal of Chemical Information and Modeling*, 49(11):2481–2488.
- [PostEra Inc.] PostEra Inc. COVID moonshot. <https://postera.ai/covid>.
- [143] Pradeepkiran, J. A., Reddy, A. P., and Reddy, P. H. (2019). Pharmacophore-based models for therapeutic drugs against phosphorylated tau in alzheimer’s disease. *Drug Discovery Today*, 24(2):616–623.
- [144] Ragoza, M., Hochuli, J., Idrobo, E., Sunseri, J., and Koes, D. R. (2017). Protein–ligand scoring with convolutional neural networks. *Journal of Chemical Information and Modeling*, 57(4):942–957.

- [145] Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., and Chen, M. (2022). Hierarchical text-conditional image generation with clip latents.
- [146] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.
- [147] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "Why should i trust you?" Explaining the predictions of any classifier. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 13-17-Aug:1135–1144.
- [148] Riniker, S. and Landrum, G. A. (2013). Open-source platform to benchmark fingerprints for ligand-based virtual screening. *Journal of Cheminformatics*, 5(1):26.
- [149] Rogers, D. and Hahn, M. (2010). Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754.
- [150] Saar, K. L., Fearon, D., Consortium, T. C. M., von Delft, F., Chodera, J. D., and Lee, A. A. (2021). Turning high-throughput structural biology into predictive inhibitor design. *bioRxiv*.
- [151] Saar, K. L., McCorkindale, W., Fearon, D., Bobby, M., Barr, H., Ben-Shmuel, A., null, London, N., von Delft, F., Chodera, J. D., Lee, A. A., Robinson, M. C., London, N., Resnick, E., Zaidmann, D., Gehrtz, P., Reddi, R. N., Gabizon, R., Barr, H., Duberstein, S., Zidane, H., Shurrush, K., Cohen, G., Solmesky, L. J., Lee, A., Jajack, A., Cvitkovic, M., Pan, J., Pai, R., Ripka, E. G., Nguyen, L., Shafeev, M., Matviiuk, T., Michurin, O., Chernyshenko, E., Bilenko, V. A., Kinakh, S. O., Logvinenko, I. G., Melnykov, K. P., Huliak, V. D., Tsurupa, I. S., Gorichko, M., Shaikh, A., Pinjari, J., Swamy, V., Pingle, M., BVNBS, S., Aimon, A., von Delft, F., Fearon, D., Dunnett, L., Douangamath, A., Dias, A., Powell, A., Neto, J. B., Skyner, R., Thompson, W., Gorrie-Stone, T., Walsh, M., Owen, D., Lukacik, P., Strain-Damerell, C., Mikolajek, H., Horrell, S., Koekemoer, L., Krojer, T., Fairhead, M., MacLean, E. M., Thompson, A., Wild, C. F., Smilova, M. D., Wright, N., von Delft, A., Gileadi, C., Rangel, V. L., Schofield, C., Salah, E., Malla, T. R., Tumber, A., John, T., Vakonakis, I., Kantsadi, A. L., Zitzmann, N., Brun, J., Kiappes, J. L., Hill, M., Witt, K. D., Alonzi, D. S., Makower, L. L., Varghese, F. S., Overheul, G. J., Miesen, P., van Rij, R. P., Jansen, J., Smeets, B., Tomésio, S., Weatherall, C., Vaschetto, M., Macdonald, H. B., Chodera, J. D., Rufa, D., Wittmann, M., Bobby, M. L., Henry, M., Glass, W. G., Eastman, P. K., Coffland, J. E., Dotson, D. L., Griffen, E. J., McCorkindale, W., Morris, A., Glen, R., Cole, J., Foster, R., Foster, H., Calmiano, M., Tennant, R. E., Heer, J., Shi, J., Jnoff, E., Hurley, M. F., Lefker, B. A., Robinson, R. P., Giroud, C., Bennett, J., Fedorov, O., Reid, S. P., Morwitzer, M. J., Cox, L., Morris, G. M., Ferla, M., Moustakas, D., Dudgeon, T., Pšenák, V., Kovar, B., Voelz, V., Carbery, A., Contini, A., Clyde, A., Ben-Shmuel, A., Sittner, A., Vitner, B. P. E. B., Bar-David, E., Tamir, H., Achdout, H., Levy, H., Glinert, I., Paran, N., Erez, N., Puni, R., Melamed, S., Weiss, S., Israely, T., Yahalom-Ronen, Y., Smalley, A., Oleinikovas, V., Spencer, J., Kenny, P. W., Ward, W., Cattermole, E., Ferrins, L., Eyermann, C. J., Milne, B. F., Godoy, A. S., Noske, G. D., Oliva, G., Fernandes, R. S., Nakamura, A. M., Gawriljuk, V. O., White, K. M., McGovern, B. L., Rosales, R., Garcia-Sastre, A., Carney, D., Chang, E., Saikatendu, K. S., Neyts, L. V. J., Donckers, K., Jochmans, D., Jonghe, S. D., Bowman, G. R., Borden, B., Singh, S., Volkamer, A., Rodriguez-Guerra, J., Fate, G., Hart, S. H., Bilenko, V. A., Kinakh, S. O., Logvinenko, I. G., Melnykov, K. P., Huliak, V. D., Tsurupa, I. S., Saar, K. L., Perry, B., Fraisse, L., Sjö, P., Boulet, P., Hahn, S., Mowbray, C., Reid,

- L., Rees, P., Huang, Q. Y. J., Zvornicanin, S. N., Shaqra, A. M., Yilmaz, N. K., Schiffer, C. A., Zhang, I., Pulido, I., Tomlinson, C., Taylor, J. C., Croll, T. I., and Brwewitz, L. (2023). Turning high-throughput structural biology into predictive inhibitor design. *Proceedings of the National Academy of Sciences*, 120(11):e2214168120.
- [152] Sacha, M., Błaż, M., Byrski, P., Włodarczyk-Pruszyński, P., and Jastrzębski, S. (2020). Molecule edit graph attention network: Modeling chemical reactions as sequences of graph edits.
- [153] Saldívar-González, F. I., Huerta-García, C. S., and Medina-Franco, J. (2020). Chemoinformatics-based enumeration of chemical libraries: a tutorial. *Journal of Cheminformatics*, 12(1):64.
- [154] Sandfort, F., Strieth-Kalthoff, F., Kühnemund, M., Beecks, C., and Glorius, F. (2020). A structure-based platform for predicting chemical reactivity. *Chem*, 6(6):1379–1390.
- [155] Santanilla, A. B., Regalado, E. L., Pereira, T., Shevlin, M., Bateman, K., Campeau, L.-C., Schneeweis, J., Berritt, S., Shi, Z.-C., Nantermet, P., Liu, Y., Helmy, R., Welch, C. J., Vachal, P., Davies, I. W., Cernak, T., and Dreher, S. D. (2015). Nanomole-scale high-throughput chemistry for the synthesis of complex molecules. *Science*, 347(6217):49–53.
- [156] Sapoval, N., Aghazadeh, A., Nute, M. G., Antunes, D. A., Balaji, A., Baraniuk, R., Barberan, C. J., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. A. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., and Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. *Nature Communications*, 13(1):1728.
- [157] Schiebel, J., Krimmer, S. G., Röwer, K., Knörlein, A., Wang, X., Park, A. Y., Stieler, M., Ehrmann, F. R., Fu, K., Radeva, N., et al. (2016). High-throughput crystallography: reliable and efficient identification of fragment hits. *Structure*, 24(8):1398–1409.
- [158] Schneider, G. (2018). Automating drug discovery. *Nature Reviews Drug Discovery*, 17(2):97–113.
- [159] Schneider, N., Lowe, D. M., Sayle, R. A., and Landrum, G. A. (2015). Development of a novel fingerprint for chemical reactions and its application to large-scale reaction classification and similarity. *Journal of Chemical Information and Modeling*, 55(1):39–53. PMID: 25541888.
- [160] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition.
- [161] Schreck, J. S., Coley, C. W., and Bishop, K. J. (2019). Learning retrosynthetic planning through simulated experience. *ACS Central Science*, 5(6):970.
- [162] Schuller, M., Correy, G. J., Gahbauer, S., Fearon, D., Wu, T., Díaz, R. E., Young, I. D., Martins, L. C., Smith, D. H., Schulze-Gahmen, U., Owens, T. W., Deshpande, I., Merz, G. E., Thwin, A. C., Biel, J. T., Peters, J. K., Moritz, M., Herrera, N., Kratochvil, H. T., null null, Aimon, A., Bennett, J. M., Neto, J. B., Cohen, A. E., Dias, A., Douangamath, A., Dunnett, L., Fedorov, O., Ferla, M. P., Fuchs, M. R., Gorrie-Stone, T. J., Holton, J. M., Johnson, M. G., Krojer, T., Meigs, G., Powell, A. J., Rack, J. G. M., Rangel, V. L., Russi, S.,

- 1 Skyner, R. E., Smith, C. A., Soares, A. S., Wierman, J. L., Zhu, K., O'Brien, P., Jura, N.,
2 Ashworth, A., Irwin, J. J., Thompson, M. C., Gestwicki, J. E., von Delft, F., Shoichet, B. K.,
3 Fraser, J. S., and Ahel, I. (2021). Fragment binding to the nsp3 macrodomain of sars-cov-2
4 identified through crystallographic screening and computational docking. *Science Advances*,
5 7(16):eabf8711.
- 6 [163] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2019a).
7 Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction.
8 *ACS Central Science*, 5(9):1572–1583.
- 9 [164] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Bekas, C., and Lee, A. A. (2019b).
10 Molecular Transformer - A Model for Uncertainty-Calibrated Chemical Reaction Prediction.
11 *ACS Central Science*, 5(9):1572–1583.
- 12 [165] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A.
13 (2019c). Molecular transformer: A model for uncertainty-calibrated chemical reaction
14 prediction. *ACS central science*, 5(9):1572–1583.
- 15 [166] Segler, M. H., Preuss, M., and Waller, M. P. (2018). Planning chemical syntheses with
16 deep neural networks and symbolic ai. *Nature*, 555(7698):604.
- 17 [167] Segler, M. H. S. (2019). World programs for model-based learning and planning in
18 compositional state and action spaces.
- 19 [168] Segler, M. H. S. and Waller, M. P. (2017). Neural-symbolic machine learning for
20 retrosynthesis and reaction prediction. *Chemistry – A European Journal*, 23(25):5966–5971.
- 21 [169] Sheridan, R. P. (2013). Time-split cross-validation as a method for estimating the good-
22 ness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4):783–
23 790.
- 24 [170] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. (2014).
25 Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine*
26 *Learning Research*, 15(56):1929–1958.
- 27 [171] Stanovsky, G., Smith, N. A., and Zettlemoyer, L. (2019). Evaluating gender bias in
28 machine translation. In *Proceedings of the 57th Annual Meeting of the Association for*
29 *Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational
30 Linguistics.
- 31 [172] Stärk, H., Ganea, O., Pattanaik, L., Barzilay, R., and Jaakkola, T. (2022). Equibind:
32 Geometric deep learning for drug binding structure prediction. In *International Conference*
33 *on Machine Learning*, pages 20503–20521. PMLR.
- 34 [173] Struble, T. J., Alvarez, J. C., Brown, S. P., Chytil, M., Cisar, J., DesJarlais, R. L.,
35 Engkvist, O., Frank, S. A., Greve, D. R., Griffin, D. J., Hou, X., Johannes, J. W., Kreatsoulas,
36 C., Lahue, B., Mathea, M., Mogk, G., Nicolaou, C. A., Palmer, A. D., Price, D. J., Robinson,
37 R. I., Salentin, S., Xing, L., Jaakkola, T., Green, W. H., Barzilay, R., Coley, C. W., and
38 Jensen, K. F. (2020). Current and future roles of artificial intelligence in medicinal chemistry
39 synthesis. *Journal of Medicinal Chemistry*, 63(16):8667–8682. PMID: 32243158.

- [174] Su, M., Yang, Q., Du, Y., Feng, G., Liu, Z., Li, Y., and Wang, R. (2019). Comparative assessment of scoring functions: The casf-2016 update. *Journal of Chemical Information and Modeling*, 59(2):895–913.
- [175] Sundararajan, M., Taly, A., and Yan, Q. (2017). Axiomatic attribution for deep networks. *34th International Conference on Machine Learning, ICML 2017*, 7:5109–5118.
- [176] Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random forest: A classification and regression tool for compound classification and qsar modeling. *Journal of Chemical Information and Computer Sciences*, 43(6):1947–1958.
- [177] Swamidass, S. J., Chen, J., Bruand, J., Phung, P., Ralaivola, L., and Baldi, P. (2005). Kernels for small molecules and the prediction of mutagenicity, toxicity and anti-cancer activity. *Bioinformatics*, 21(suppl_1):359–368.
- [178] Temml, V., Voss, C. V., Dirsch, V. M., and Schuster, D. (2014). Discovery of new liver x receptor agonists by pharmacophore modeling and shape-based virtual screening. *Journal of Chemical Information and Modeling*, 54(2):367–371.
- [179] Tetko, I. V. (2002). Neural network studies. 4. introduction to associative neural networks. *Journal of Chemical Information and Computer Sciences*, 42(3):717–728. PMID: 12086534.
- [180] Tetko, I. V., Karpov, P., Van Deursen, R., and Godin, G. (2020). State-of-the-art augmented nlp transformer models for direct and single-step retrosynthesis. *Nature communications*, 11(1):1–11.
- [181] Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O., and Bjerrum, E. J. (2020). Datasets and their influence on the development of computer assisted synthesis planning tools in the pharmaceutical domain. *Chem. Sci.*, 11:154–168.
- [182] The COVID Moonshot Consortium (2020). Covid moonshot: open science discovery of sars-cov-2 main protease inhibitors by combining crowdsourcing, high-throughput experiments, computational simulations, and machine learning. *bioRxiv*, doi:10.1101/2020.10.29.339317.
- [183] The COVID Moonshot Consortium, Achdout, H., Aimon, A., Bar-David, E., Barr, H., Ben-Shmuel, A., Bennett, J., Bilenko, V. A., Bilenko, V. A., Boby, M. L., Borden, B., Bowman, G. R., Brun, J., BVNBS, S., Calmiano, M., Carbery, A., Carney, D., Cattermole, E., Chang, E., Chernyshenko, E., Chodera, J. D., Clyde, A., Coffland, J. E., Cohen, G., Cole, J., Contini, A., Cox, L., Cvitkovic, M., Dias, A., Donckers, K., Dotson, D. L., Douangamath, A., Duberstein, S., Dudgeon, T., Dunnett, L., Eastman, P. K., Erez, N., Eyermann, C. J., Fairhead, M., Fate, G., Fearon, D., Fedorov, O., Ferla, M., Fernandes, R. S., Ferrins, L., Foster, R., Foster, H., Gabizon, R., Garcia-Sastre, A., Gawriljuk, V. O., Gehrtz, P., Gileadi, C., Giroud, C., Glass, W. G., Glen, R., Glinert, I., Godoy, A. S., Gorichko, M., Gorrie-Stone, T., Griffen, E. J., Hart, S. H., Heer, J., Henry, M., Hill, M., Horrell, S., Huliak, V. D., Hurley, M. F., Israely, T., Jajack, A., Jansen, J., Jnoff, E., Jochmans, D., John, T., Jonghe, S. D., Kantsadi, A. L., Kenny, P. W., Kiappes, J. L., Kinakh, S. O., Koekemoer, L., Kovar, B., Krojer, T., Lee, A., Lefker, B. A., Levy, H., Logvinenko, I. G., London, N., Lukacik, P., Macdonald, H. B., MacLean, B., Malla, T. R., Matviuk, T., McCorkindale, W., McGovern, B. L., Melamed, S., Melnykov, K. P., Michurin, O., Mikolajek, H., Milne, B. F., Morris, A., Morris, G. M., Morwitzer, M. J., Moustakas, D., Nakamura, A. M., Neto, J. B., Neyts, J.,

Nguyen, L., Noske, G. D., Oleinikovas, V., Oliva, G., Overheul, G. J., Owen, D., Pai, R., Pan, J., Paran, N., Perry, B., Pingle, M., Pinjari, J., Politi, B., Powell, A., Psenak, V., Puni, R., Rangel, V. L., Reddi, R. N., Reid, S. P., Resnick, E., Ripka, E. G., Robinson, M. C., Robinson, R. P., Rodriguez-Guerra, J., Rosales, R., Rufa, D., Saar, K., Saikatendu, K. S., Schofield, C., Shafeev, M., Shaikh, A., Shi, J., Shurrush, K., Singh, S., Sittner, A., Skyner, R., Smalley, A., Smeets, B., Smilova, M. D., Solmesky, L. J., Spencer, J., Strain-Damerell, C., Swamy, V., Tamir, H., Tennant, R., Thompson, W., Thompson, A., Tomasio, S., Tsurupa, I. S., Tumber, A., Vakonakis, I., van Rij, R. P., Vangeel, L., Varghese, F. S., Vaschetto, M., Vitner, E. B., Voelz, V., Volkamer, A., von Delft, F., von Delft, A., Walsh, M., Ward, W., Weatherall, C., Weiss, S., White, K. M., Wild, C. F., Wittmann, M., Wright, N., Yahalom-Ronen, Y., Zaidmann, D., Zidane, H., and Zitzmann, N. (2022). Open science discovery of oral non-covalent sars-cov-2 main protease inhibitor therapeutics. *bioRxiv*.

[184] Todeschini, R., Consonni, V., Xiang, H., Holliday, J., Buscema, M., and Willett, P. (2012). Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11):2884–2901. PMID: 23078167.

[185] Trnka, T. M. and Grubbs, R. H. (2001). The development of 12x2ruchr olefin metathesis catalysts: An organometallic success story. *Accounts of Chemical Research*, 34(1):18–29. PMID: 11170353.

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

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

[188] Vandenberk, J., Kennis, L. E. J., Van Heertum, A. H. M. T., and Van der Aa, M. J. M. C. (1981). 1,3-dihydro-1-[(1-piperidinyl)alkyl]-2h-benzimidazol-2-one derivatives.

[189] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 2017-Decem(Nips):5999–6009.

[190] Verdonk, M. L., Cole, J. C., Hartshorn, M. J., Murray, C. W., and Taylor, R. D. (2003). Improved protein–ligand docking using gold. *Proteins: Structure, Function, and Bioinformatics*, 52(4):609–623.

[191] Volkamer, A., Riniker, S., Nittinger, E., Lanini, J., Grisoni, F., Evertsson, E., Rodríguez-Pérez, R., and Schneider, N. (2023). Machine learning for small molecule drug discovery in academia and industry. *Artificial Intelligence in the Life Sciences*, 3:100056.

[192] Vuorinen, A. and Schuster, D. (2015). Methods for generating and applying pharmacophore models as virtual screening filters and for bioactivity profiling. *Methods*, 71:113–134.

[193] Walters, W., Stahl, M. T., and Murcko, M. A. (1998). Virtual screening—an overview. *Drug Discovery Today*, 3(4):160–178.

- [194] Walters, W. P. (2019). Virtual chemical libraries. *Journal of Medicinal Chemistry*, 62(3):1116–1124. 1 2
- [195] Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The pdbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *Journal of Medicinal Chemistry*, 47(12):2977–2980. 3 4 5
- [196] Wei, J. N., Duvenaud, D., and Aspuru-Guzik, A. (2016). Neural networks for the prediction of organic chemistry reactions. *ACS Central Science*, 2(10):725–732. 6 7
- [197] Weininger, D. (1988). SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36. 8 9 10
- [198] Weininger, D., Weininger, A., and Weininger, J. L. (1989). SMILES. 2. Algorithm for Generation of Unique SMILES Notation. *Journal of Chemical Information and Computer Sciences*, 29(2):97–101. 11 12 13
- [199] Willett, P., Barnard, J. M., and Downs, G. M. (1998). Chemical similarity searching. *Journal of Chemical Information and Computer Sciences*, 38(6):983–996. 14 15
- [200] Wong, F., Krishnan, A., Zheng, E. J., Stärk, H., Manson, A. L., Earl, A. M., Jaakkola, T., and Collins, J. J. (2022). Benchmarking alphafold-enabled molecular docking predictions for antibiotic discovery. *Molecular Systems Biology*, 18(9):e11081. 16 17 18
- [201] Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. (2018). Moleculenet: a benchmark for molecular machine learning. *Chem. Sci.*, 9:513–530. 19 20 21
- [202] Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., and Barzilay, R. (2019). Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59(8):3370–3388. PMID: 31361484. 22 23 24 25
- [203] Yang, Y., Zheng, S., Su, S., Zhao, C., Xu, J., and Chen, H. (2020). Syntalinker: automatic fragment linking with deep conditional transformer neural networks. *Chem. Sci.*, 11:8312–8322. 26 27 28
- [204] Yu, H. S., Modugula, K., Ichihara, O., Kramschuster, K., Keng, S., Abel, R., and Wang, L. (2021). General theory of fragment linking in molecular design: Why fragment linking rarely succeeds and how to improve outcomes. *Journal of Chemical Theory and Computation*, 17(1):450–462. PMID: 33372778. 29 30 31 32

