

## Precision Fragment Addition: Domain-Specific DeepFrag2 Models for Smarter Lead Optimization (Supporting Information)

César R. García-Jacas<sup>1</sup>, Harrison Green<sup>1</sup>, Shayne D. Wierbowski<sup>2</sup>, Jacob D. Durrant<sup>1,\*</sup>

<sup>1</sup> Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA, USA

<sup>2</sup> Machine Learning and Computational Sciences, Pfizer Research and Development, Cambridge, MA, USA

\* To whom correspondence should be addressed.

### Strengths and Weaknesses of Top-K Metric

In the current work, we used Top-K accuracy as our primary evaluation metric because it provides an easy-to-interpret approach for comparing model performance, but the method certainly has its limitations. For example, the “correct” fragment, as defined by the examples in our dataset, is certainly not the absolute optimal choice in all scenarios. We encountered a good illustration of this limitation when testing DeepFrag on the FDA-approved drug oseltamivir bound to influenza neuraminidase (2HU4:A<sup>1</sup>). When we removed the oseltamivir carboxylate and used DeepFrag to predict a suitable replacement, its top selection was a phosphonate group, not the “correct” carboxylate group. And yet, though no crystal structure including oseltamivir phosphonic acid is present in the Protein Data Bank, studies suggest that phosphonate is a suitable bioisostere of carboxylate in this case; indeed, the phosphonic acid form is more active than the “correct” carboxylate form (IC<sub>50</sub>: 0.52 nM vs. 0.73 nM)<sup>2</sup>.

Top-K accuracy is also limited as a metric because it treats all incorrect predictions equally without considering the chemical similarity or biological equivalence of the suggested fragments. In an ideal assessment, predictions that are chemically similar to a ground-truth fragment should not be penalized as much as entirely unrelated predictions. See Figure S1, second row, for an example of a prediction that is technically wrong and yet clearly very similar to the ground truth.

A third limitation pertains specifically to Top-1 accuracy (i.e., how often the most-similar label-set fragment is identical to the ground-truth fragment). In real-world drug discovery, medicinal chemists typically synthesize and evaluate a series of compounds rather than a single (top) candidate. A model’s ability to suggest multiple promising fragments is often more valuable than its ability to predict a single “correct” answer. We therefore consider Top-8 accuracy (i.e., how often the ground-truth fragment is among the eight most-similar label-set fragments) to be a more appropriate metric for assessing DeepFrag’s performance.

## 2048-Bit RDKfingerprints Enable Improved Top-K Accuracy

We encoded molecular fragments as 2048-bit RDKfingerprints for DeepFrag2 training and label-set lookup (see main text). To justify this fingerprint size, we calculated 2048-bit RDKfingerprints for all unique fragments in the all-fragment dataset. We found that each of the 2048 bits was occupied in at least some of the fragment fingerprints (min prevalence 3.7%, max prevalence 69.7%, average prevalence 10.2%), demonstrating that no bit went unused.

Further, we found that RDKfingerprints performed slightly better than other molecular fingerprints tested. For the sake of comparison, we trained three versions of the general (all-fragments) model for sixty epochs using 2048-bit RDKfingerprints, 2048-bit circular Morgan fingerprints<sup>3</sup>, and 166-bit structural-key MACCS fingerprints<sup>4</sup>, respectively, using the same all-fragment training set described in the main text. We then evaluated each last-epoch trained model on the same main-text testing set and calculated Top-K metrics. RDKfingerprint DeepFrag2 consistently had Top-K metrics slightly better than Morgan and MACCS DeepFrag2 (Table S1), though Morgan DeepFrag2 was comparable.

Table S1. Top-K metrics for DeepFrag2 models trained using RDKfingerprints, Morgan fingerprints, and MACCS fingerprints.

<b>Fingerprint</b>	<b>Size</b>	<b>Top-1</b>	<b>Top-8</b>	<b>Top-16</b>
RDKfingerprint	2048	44.4	56.1	59.0
Morgan	2048	44.0	55.4	58.0
MACCS	166	42.1	52.3	55.3

“Fingerprint” indicates the molecular fingerprint used, and “Size” indicates the length of the fingerprint vector. Each model was trained for 60 epochs on the same all-fragment training set used to create the general model described in the main text. Each final-epoch model was applied to the same general-model testing set, and the testing-set predictions were used to calculate Top-K metrics (K = 1, 8, 16; i.e., the percentage of cases where the correct fragment was among the top *K* predictions).

## Complete Report of Top-K Metrics for Each DeepFrag2 Model

In the main text, we report the Top-1 and Top-8 metrics for each trained DeepFrag2 model. Table S2 extends this information to include Top-16, Top-32, and Top-64 metrics.

Table S2. Performance comparison of targeted and general DeepFrag models for various fragment types.

Fragments	Targeted					General					Random					Label-Set Size
	T1	T8	T16	T32	T64	T1	T8	T16	T32	T64	T1	T8	T16	T32	T64	
<i>Type (Size)</i>																
all (any)	N/A	N/A	N/A	N/A	N/A	44.4	56.1	59.0	62.3	66.1	0.04	0.3	0.7	1.3	2.6	2,453
all ( $\leq 3$ )	69.9	86.2	89.2	92.3	97.9	67.1	82.2	86.5	90.5	97.5	1.3	10.1	20.3	40.5	81.0	79
all ( $\geq 4$ )	17.5	30.2	34.0	37.7	42.4	16.3	28.6	32.1	36.3	42.0	0.04	0.3	0.7	1.3	2.7	2,374
aromatic ( $\geq 4$ )	14.4	25.1	28.9	32.9	39.5	11.2	19.1	22.9	26.5	32.4	0.1	0.8	1.6	3.3	6.6	975
aliphatic ( $\geq 4$ )	23.6	41.6	46.4	52.6	58.8	20.7	37.1	41.8	47.1	54.3	0.1	0.6	1.1	2.3	4.6	1,399
acid ( $\geq 4$ )	50.0	64.7	68.8	73.8	79.5	44.4	61.8	66.6	72.7	78.4	0.4	3.0	6.0	12.0	24.0	267
base ( $\geq 4$ )	20.1	37.9	49.5	58.0	66.4	13.9	30.0	37.7	46.2	58.8	0.3	2.2	4.4	8.8	17.6	363

“Fragments” indicates the chemical properties (“Type”) and heavy atom counts (“Size”) of the fragments used to train the corresponding model. “Strucs” refers to the number of protein-ligand structures, and “Examples” refers to the number of protein/trimmed-ligand/fragment examples derived from those structures. The targeted models were trained only on examples whose fragments had the properties described in the “Fragments” column. The general model was trained on all example types regardless of fragment properties. “Random” indicates a random baseline (i.e., the results obtained when selecting compounds at random from the corresponding test-set-derived label set). Label-Set Size indicates the number of unique fragments in the testing-set-derived label set. Top-K values ( $K = 1, 8, 16, 32, 64$ ) represent the percentage of cases where the correct fragment was among the top  $K$  predictions when the label set comprised all the fragments of the respective targeted (property-specific) testing set.

## Examples of General-Model Performance

	Ground Truth	Top 5 Predictions				
Training: All Label Set: All Top-8: 56.1%		 0.679	 0.677	 0.675	 0.664	 0.660
		 0.860	 0.621	 0.573	 0.516	 0.506
		 0.529	 0.525	 0.518	 0.516	 0.514
		 1.000	 0.178	 0.151	 0.126	 0.126
		 0.956	 0.689	 0.644	 0.603	 0.600

Figure S1. General DeepFrag2 model (trained on fragments of all size classes and chemical properties) applied to five randomly selected testing-set examples. The label set included all testing-set fragments. In many cases, even when DeepFrag2 doesn't select the correct fragment, it selects chemically similar fragments (e.g., row 2).

## Binned Top-1 accuracy analysis as a function of cosine similarity

We expect many DeepFrag users will ask, "Given a set of top DeepFrag-selected fragments, what are the chances a useful fragment is among them?" To this end, we considered Top-1 cosine similarity scores, defined as the cosine similarity between the Top-1 (most similar) label-set fragment fingerprint and the corresponding DeepFrag-predicted fingerprint. We examined whether there exists a reasonable Top-1 cosine-similarity threshold above which one can reliably conclude that the ground-truth fragment is among the most similar label-set compounds. If so, medicinal chemists could use Top-1 cosine similarity as a practical confidence metric.

For each model, we first determined the Top-1 cosine-similarity score of each example in the corresponding testing set. We then binned the testing-set examples by their Top-1 cosine similarity scores and calculated the Top-1 accuracy within each bin (Figure S2, purple bars). In other words, for each Top-1 cosine-similarity bin, we determined the percentage of cases in which the ground-truth fragment was in fact the top-ranked selection.

We extended this analysis to include Top-8 and Top-16 accuracy (Figure S2, gray and black lines, respectively). Here we again binned the examples by the same Top-1 cosine similarity scores, but we considered how often the ground-truth fragment was among the Top-8 (and Top-16) most similar fragments from the respective label set.

Our results revealed a strong relationship between high Top-1 cosine similarity and increased Top-1, Top-8, and Top-16 accuracy, with the highest-similarity bins demonstrating exceptional accuracy in many cases (Figures S2-S7). This finding suggests that high Top-1 cosine similarity can

indeed be interpreted as a confidence score, providing a quantitative measure of prediction reliability. If the cosine similarity associated with the Top-1 selected fragment is greater than 0.975, the selected fragment is likely to match the ground-truth fragment.

That said, in practice users may wish to select a lower threshold. Fragments with somewhat lower cosine-similarity scores could nevertheless be structurally and chemically relevant bioisosteres, making them effective fragment additions.

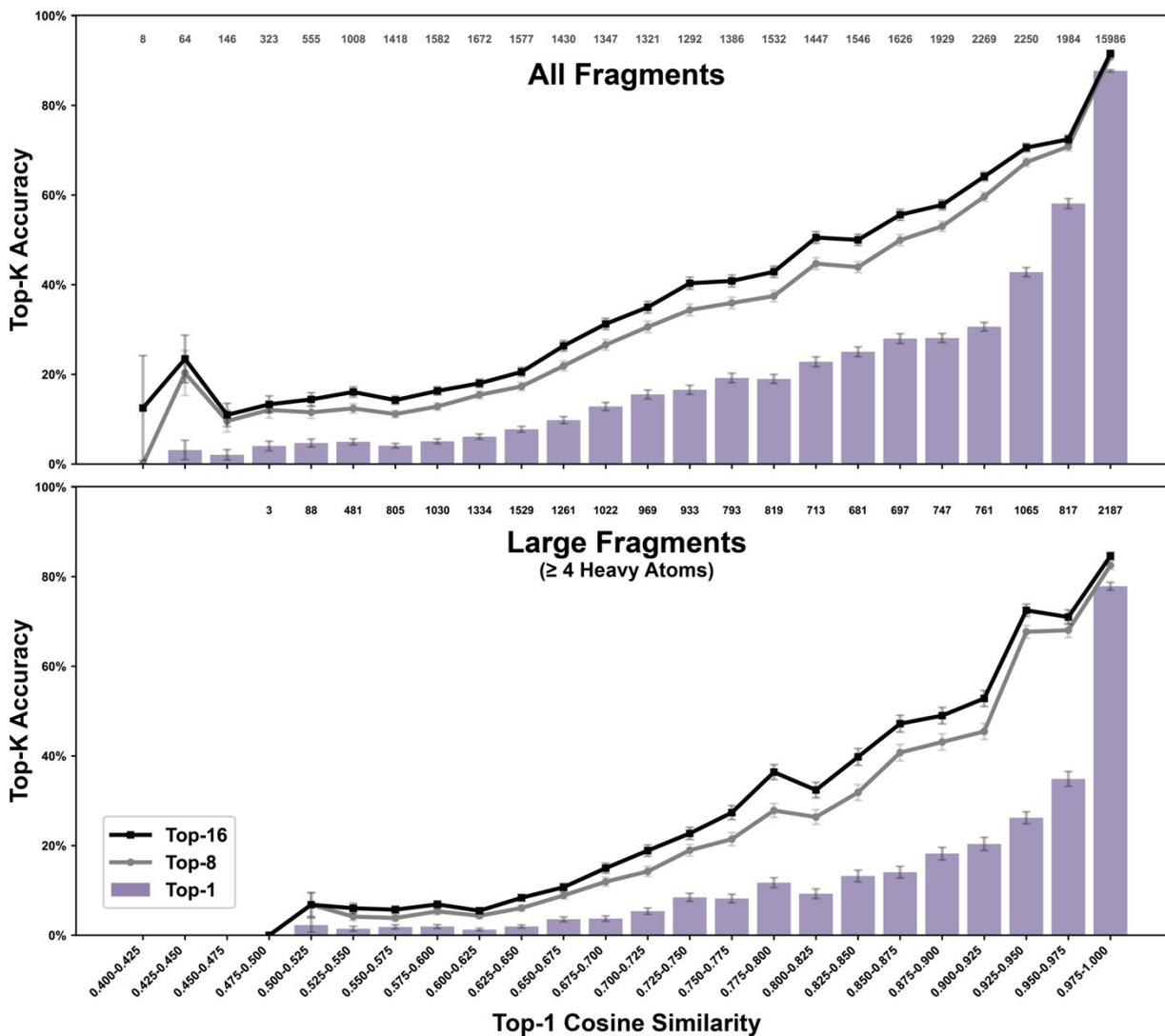


Figure S2. Relationship between Top-1 cosine similarity and Top-K accuracy. The plots show Top-1 (purple bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. The upper panel shows results from the general model (trained on fragments of all chemical properties and size classes), and the lower panel shows results from the large-fragment model ( $\geq 4$  heavy atoms). In both cases, the label set was comprised of all fragments derived from the corresponding testing set. Numbers above each bin indicate the sample size. Error bars represent standard error.

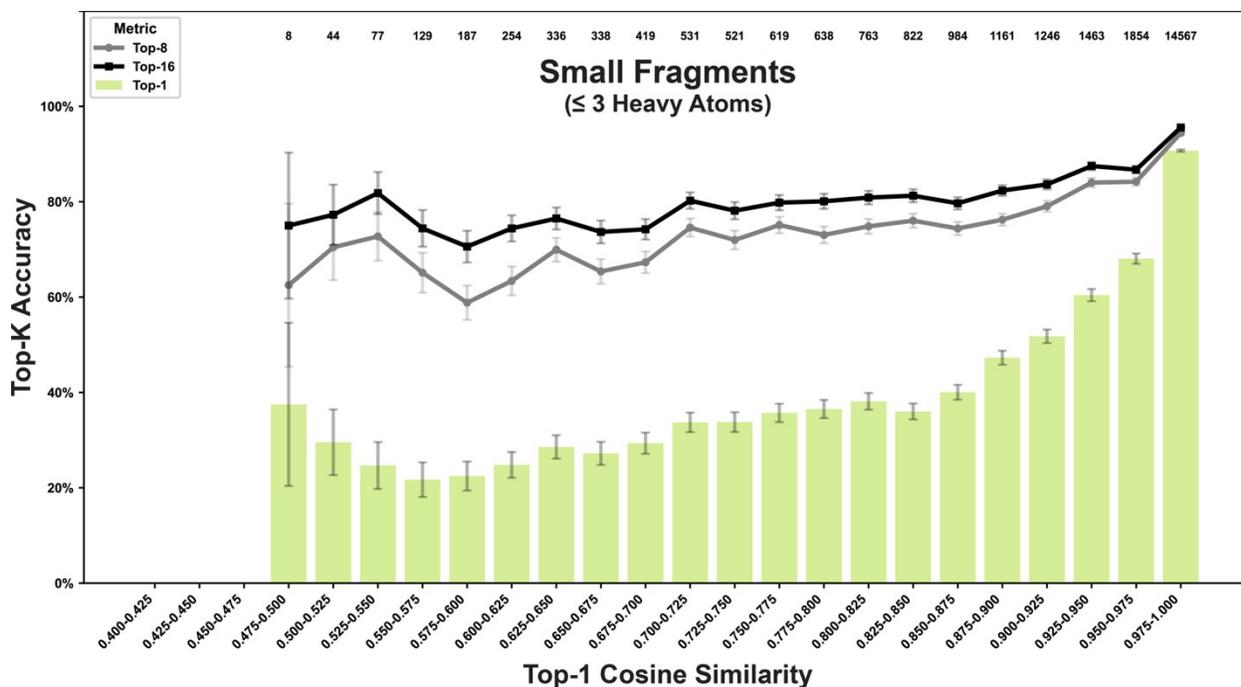


Figure S3. Relationship between Top-1 cosine similarity and Top-K accuracy for the small-fragment model ( $\leq 3$  heavy atoms). The plots show Top-1 (bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. Numbers above each bin indicate the sample size. Error bars represent standard error.

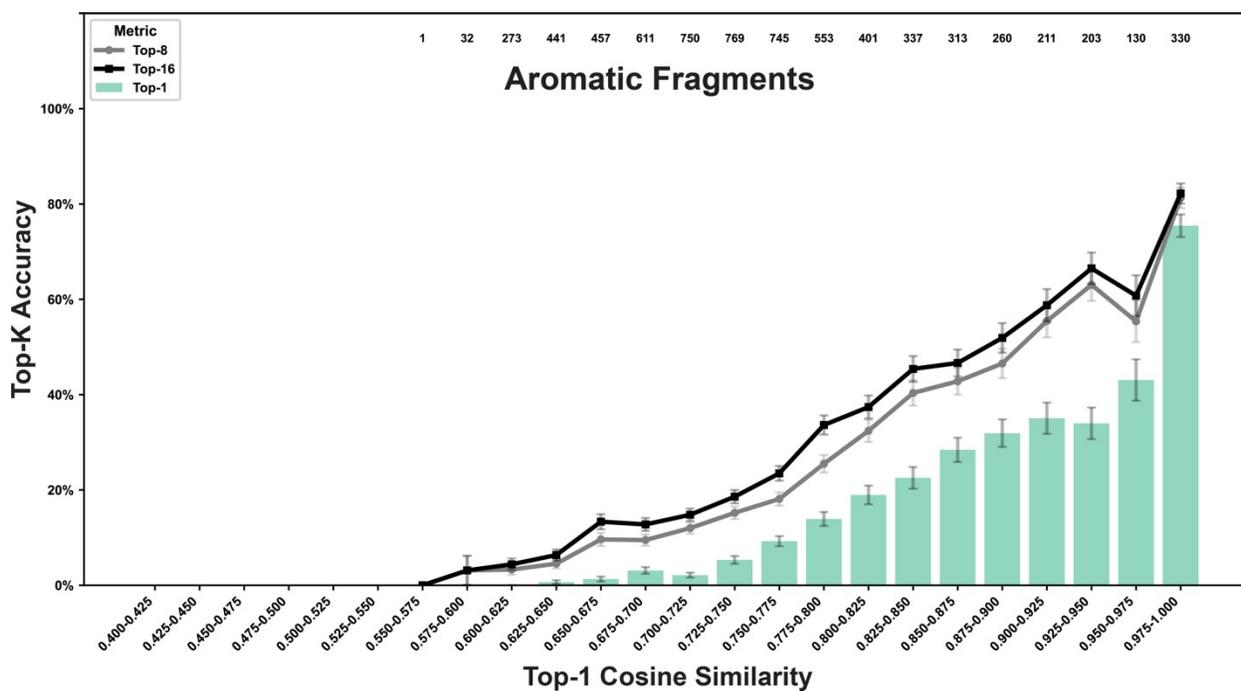


Figure S4. Relationship between Top-1 cosine similarity and Top-K accuracy for the aromatic-fragment model ( $\geq 4$  heavy atoms). The plots show Top-1 (bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. Numbers above each bin indicate the sample size. Error bars represent standard error.

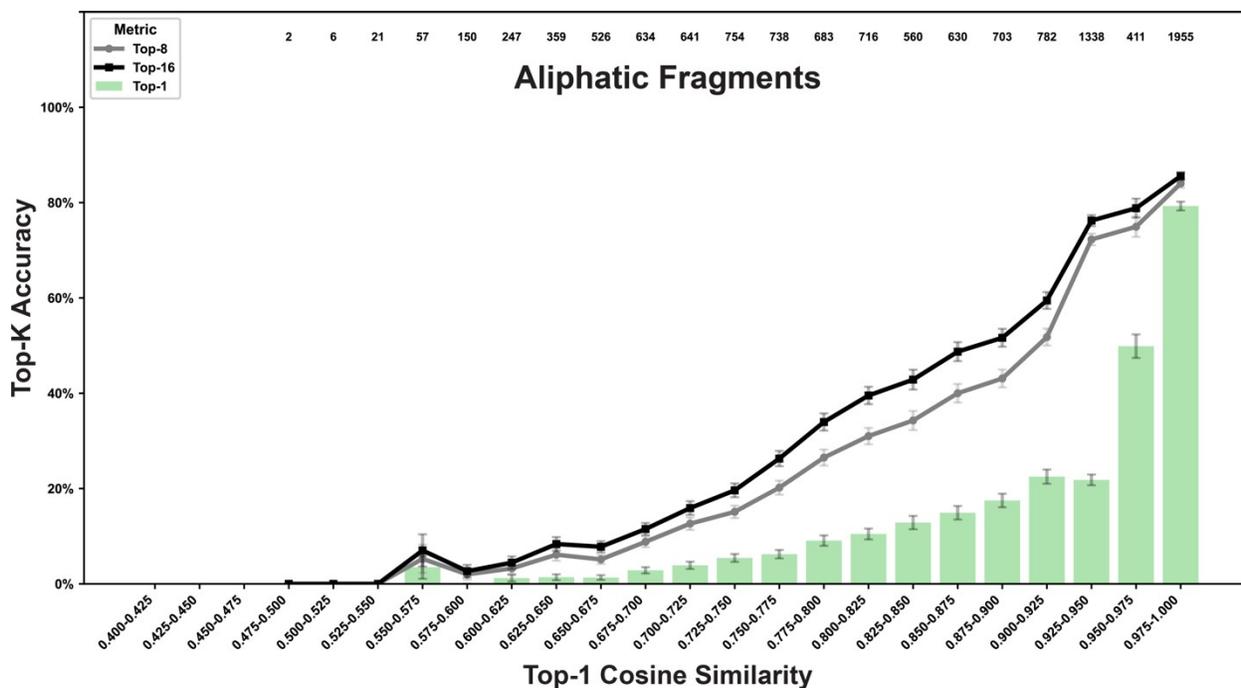


Figure S5. Relationship between Top-1 cosine similarity and Top-K accuracy for the aliphatic-fragment model ( $\geq 4$  heavy atoms). The plots show Top-1 (bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. Numbers above each bin indicate the sample size. Error bars represent standard error.

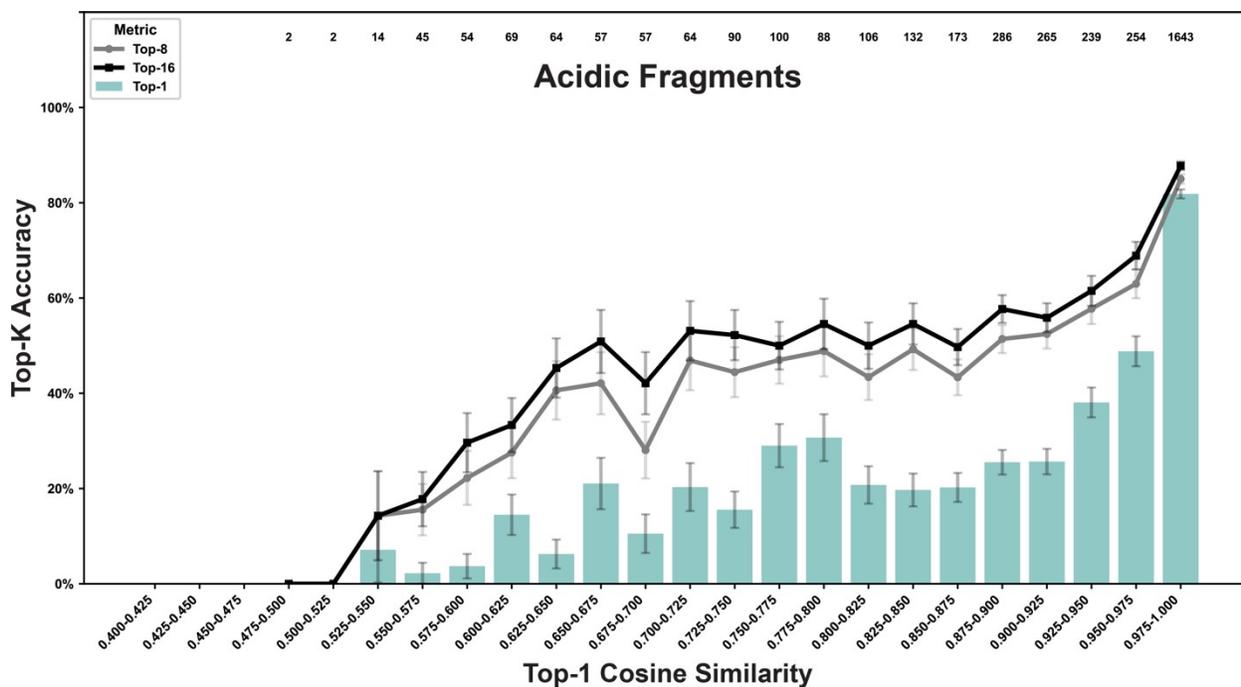


Figure S6. Relationship between Top-1 cosine similarity and Top-K accuracy for the acidic-fragment model ( $\geq 4$  heavy atoms). The plots show Top-1 (bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. Numbers above each bin indicate the sample size. Error bars represent standard error.

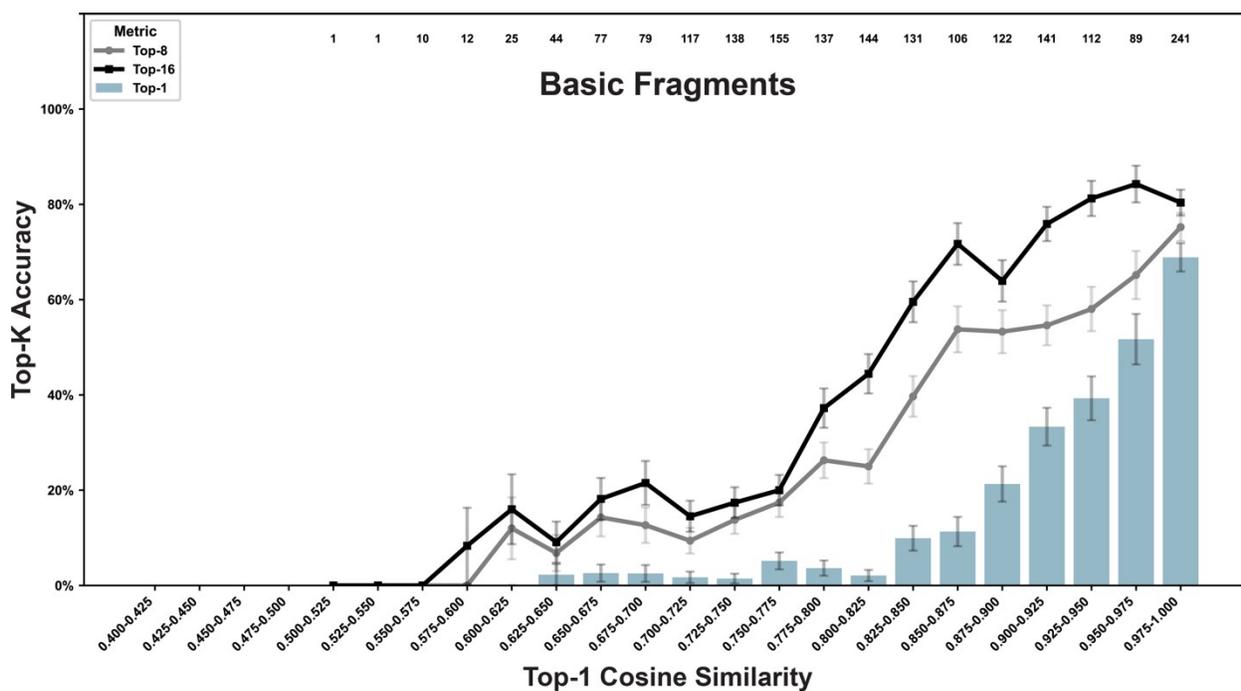


Figure S7. Relationship between Top-1 cosine similarity and Top-K accuracy for the basic-fragment model ( $\geq 4$  heavy atoms). The plots show Top-1 (bars), Top-8 (gray line), and Top-16 (black line) accuracies binned by Top-1 cosine similarity. Numbers above each bin indicate the sample size. Error bars represent standard error.

Accuracy of fragment-specific predictions as a function of cosine similarity

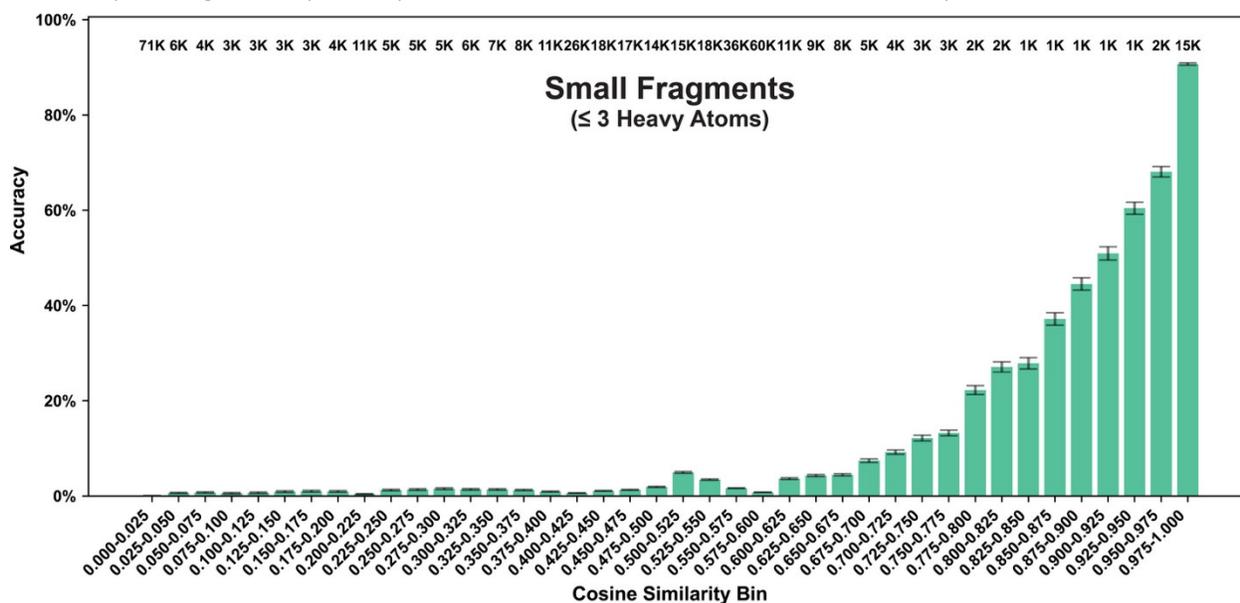


Figure S8. Relationship between a fragment's cosine-similarity score and its likelihood of matching the corresponding ground-truth fragment, when using the small-fragment model ( $\leq 3$  heavy atoms) and a label set comprised of all small fragments from the testing set. Source data includes all Top-16 fragment selections from all small-fragment testing-set examples. The plots show accuracy (bars) binned by cosine similarity between DeepFrag-predicted and label-set fingerprints, where accuracy represents the percentage of cases in which the recommended fragment matched the ground-truth fragment within each bin. Numbers above each bin indicate the sample size, abbreviated where necessary (K represents thousands). Error bars represent standard error.

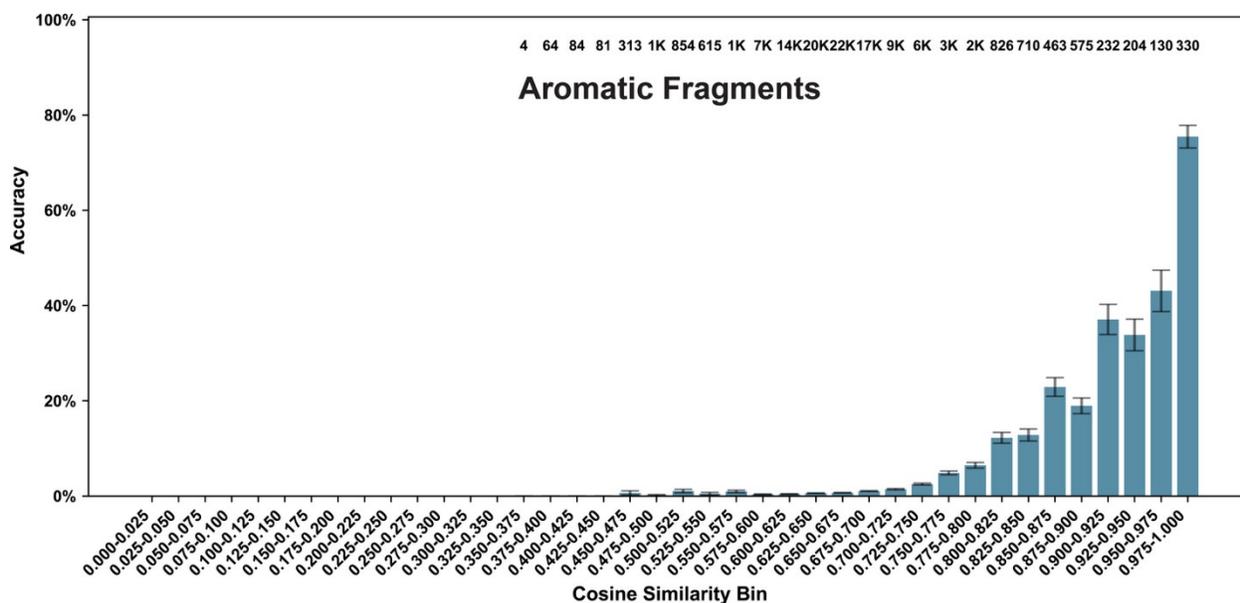


Figure S9. Relationship between a fragment's cosine-similarity score and its likelihood of matching the corresponding ground-truth fragment, when using the aromatic-fragment model ( $\geq 4$  heavy atoms) and a label set comprised of all aromatic fragments from the testing set. Source data includes all Top-16 fragment selections from all aromatic-fragment testing-set examples. The plots show accuracy (bars) binned by cosine similarity between DeepFrag-predicted and label-set fingerprints, where accuracy represents the percentage of cases in which the recommended fragment matched the ground-truth fragment within each bin. Numbers above each bin indicate the sample size, abbreviated where necessary (K represents thousands). Error bars represent standard error.

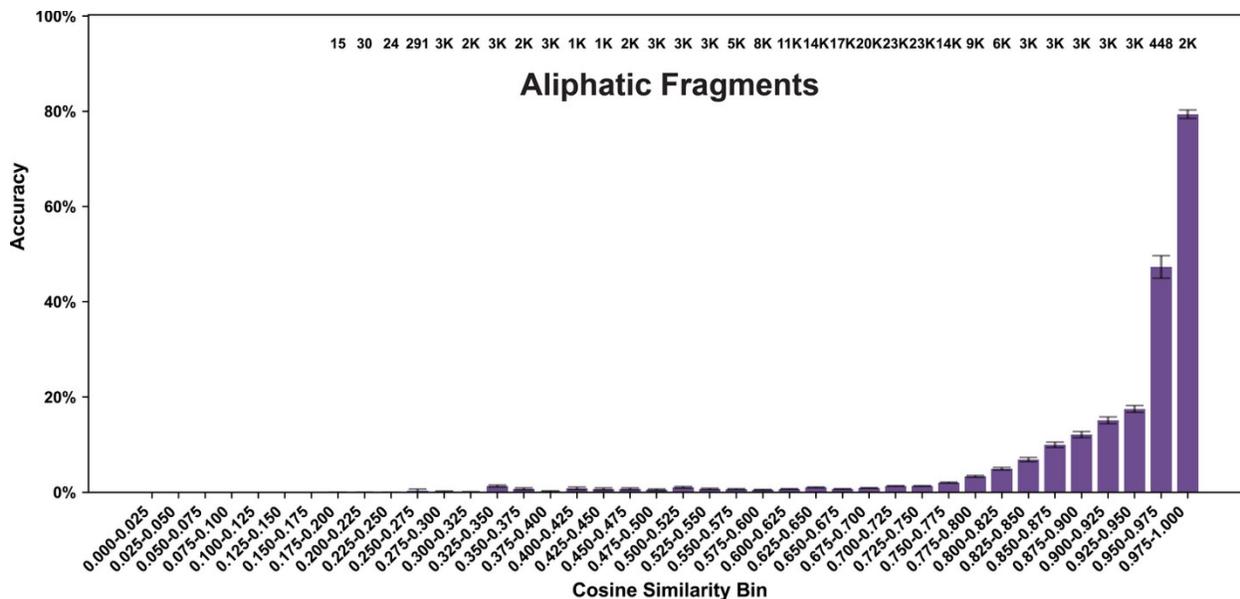


Figure S10. Relationship between a fragment's cosine-similarity score and its likelihood of matching the corresponding ground-truth fragment, when using the aliphatic-fragment model ( $\geq 4$  heavy atoms) and a label set comprised of all aliphatic fragments from the testing set. Source data includes all Top-16 fragment selections from all aliphatic-fragment testing-set examples. The plots show accuracy (bars) binned by cosine similarity between DeepFrag-predicted and label-set fingerprints, where accuracy represents the percentage of cases in which the recommended fragment matched the ground-truth fragment within each bin. Numbers above each bin indicate the sample size, abbreviated where necessary (K represents thousands). Error bars represent standard error.

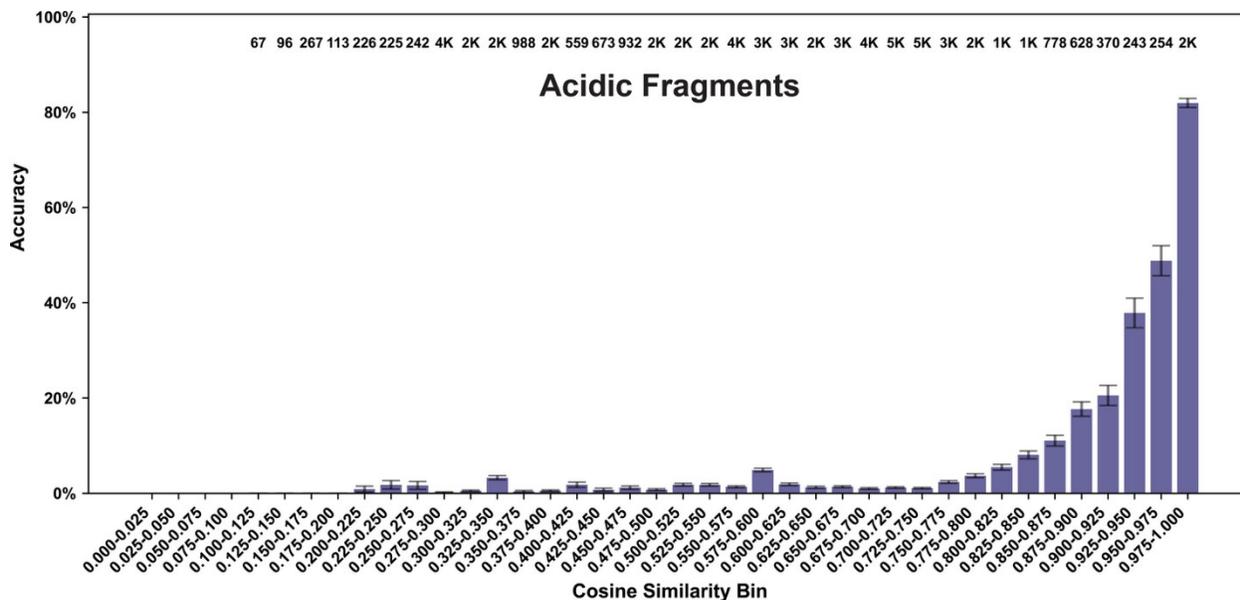


Figure S11. Relationship between a fragment's cosine-similarity score and its likelihood of matching the corresponding ground-truth fragment, when using the acidic-fragment model ( $\geq 4$  heavy atoms) and a label set comprised of all acidic fragments from the testing set. Source data includes all Top-16 fragment selections from all acidic-fragment testing-set examples. The plots show accuracy (bars) binned by cosine similarity between DeepFrag-predicted and label-set fingerprints, where accuracy represents the percentage of cases in which the recommended fragment matched the ground-truth fragment within each bin. Numbers above each bin indicate the sample size, abbreviated where necessary (K represents thousands). Error bars represent standard error.

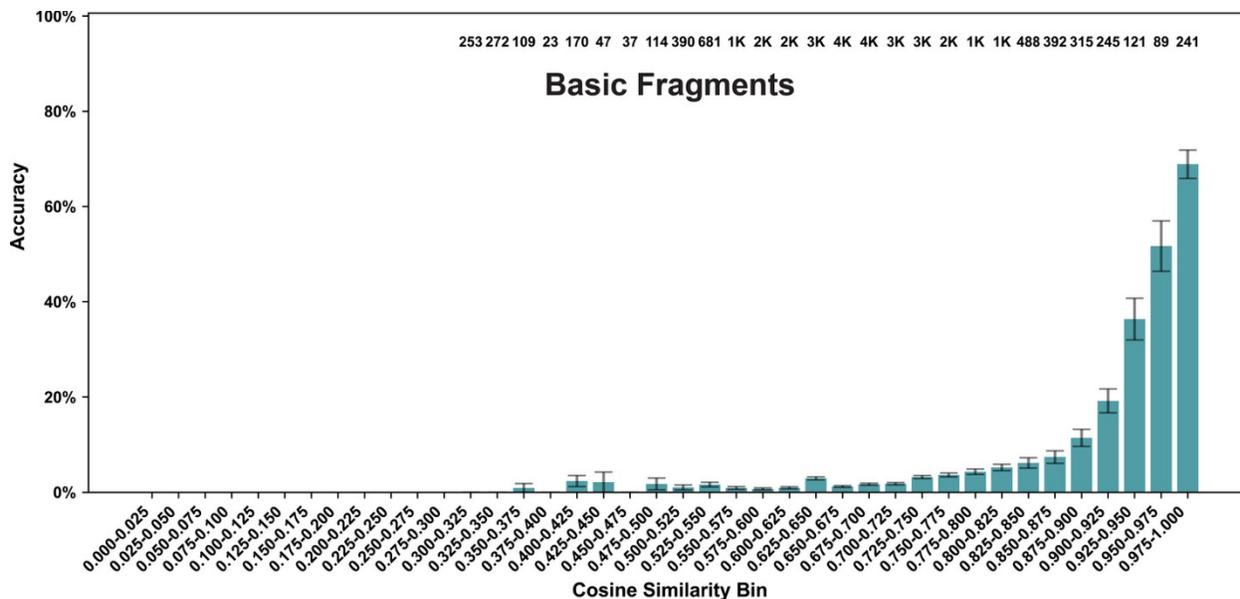


Figure S12. Relationship between a fragment's cosine-similarity score and its likelihood of matching the corresponding ground-truth fragment, when using the basic-fragment model ( $\geq 4$  heavy atoms) and a label set comprised of all basic fragments from the testing set. Source data includes all Top-16 fragment selections from all basic-fragment testing-set examples. The plots show accuracy (bars) binned by cosine similarity between DeepFrag-predicted and label-set fingerprints, where accuracy represents the percentage of cases in which the recommended fragment matched the ground-truth fragment within each bin. Numbers above each bin indicate the sample size, abbreviated where necessary (K represents thousands). Error bars represent standard error.

Protein-specific finetuned DeepFrag models

	Ground Truth	Top 5 Predictions				
Training: BACE $\geq 4$ Label Set: BACE $\geq 4$ Top-8: 47.0%		 0.906	 0.651	 0.549	 0.520	 0.447
		 0.668	 0.651	 0.651	 0.620	 0.587
		 0.852	 0.732	 0.730	 0.713	 0.712
		 0.595	 0.570	 0.558	 0.552	 0.549
		 0.677	 0.657	 0.652	 0.622	 0.589
Training: All $\geq 4$ , Epoch 60 Label Set: BACE $\geq 4$ Top-8: 24.1%		 0.474	 0.461	 0.445	 0.434	 0.430
		 0.507	 0.416	 0.319	 0.283	 0.237
		 0.594	 0.552	 0.537	 0.535	 0.529
		 0.544	 0.540	 0.529	 0.524	 0.524
		 0.498	 0.409	 0.314	 0.278	 0.234

Figure S13. BACE fine-tuning. The upper five rows show randomly selected testing-set predictions after 30 generations of fine-tuning on beta-secretase 1 (BACE) protein/ligand complexes, with corresponding ground-truth fragments in the left most column. The fine-tuned model was trained to predict large fragments ( $\geq 4$  heavy atoms). The bottom five rows show the predictions of the model before finetuning when applied to the same examples. This foundational model was trained on large fragments (independent of chemical property or protein target). In both cases, the label set includes only large fragments, derived from the BACE testing-set examples.

	Ground Truth	Top 5 Predictions				
Training: All $\geq 4$ , Epoch 60		 0.857	 0.553	 0.480	 0.392	 <u>0.389</u>
		 <u>0.628</u>	 0.599	 0.566	 0.529	 0.506
		 0.729	 0.620	 <u>0.614</u>	 0.582	 0.567
		 <u>1.000</u>	 0.808	 0.742	 0.564	 0.501
		 <u>0.985</u>	 0.552	 0.430	 0.118	 0.109
Training: GluR2 $\geq 4$		 0.549	 0.541	 <u>0.540</u>	 0.528	 0.482
		 0.496	 0.471	 0.458	 0.423	 0.415
		 0.587	 0.575	 0.572	 0.546	 0.538
		 <u>0.891</u>	 0.828	 0.819	 0.724	 0.721
		 0.538	 0.530	 0.475	 0.470	 0.460
Label Set: GluR2 $\geq 4$		 0.729	 0.620	 <u>0.614</u>	 0.582	 0.567
		 <u>1.000</u>	 0.808	 0.742	 0.564	 0.501
		 <u>0.628</u>	 0.599	 0.566	 0.529	 0.506
		 0.857	 0.553	 0.480	 0.392	 <u>0.389</u>
		 0.729	 0.620	 <u>0.614</u>	 0.582	 0.567

Figure S14. GluR2 fine-tuning. The upper five rows show randomly selected testing-set predictions after 30 generations of fine-tuning on glutamate receptor (GluR2) protein/ligand complexes, with corresponding ground-truth fragments in the left most column. The fine-tuned model was trained to predict large fragments ( $\geq 4$  heavy atoms). The bottom five rows show the predictions of the model before finetuning when applied to the same examples. This foundational model was trained on large fragments (independent of chemical property or protein target). In both cases, the label set includes only large fragments, derived from the GluR2 testing-set examples.

	Ground Truth	Top 5 Predictions				
Training: All $\geq 4$ , Epoch 60		 0.999	 0.686	 0.569	 0.523	 0.449
		 1.000	 0.749	 0.743	 0.693	 0.652
		 0.890	 0.759	 0.742	 0.659	 0.603
		 0.732	 0.674	 0.650	 0.645	 0.642
		 0.834	 0.768	 0.671	 0.602	 0.584
Label Set: HIV-PR $\geq 4$		 0.878	 0.768	 0.653	 0.646	 0.637
		 0.857	 0.716	 0.680	 0.662	 0.655
		 0.660	 0.615	 0.607	 0.602	 0.589
		 0.582	 0.577	 0.569	 0.554	 0.552
		 0.545	 0.543	 0.541	 0.529	 0.528

Figure S15. HIV-PR fine-tuning. The upper five rows show randomly selected testing-set predictions after 30 generations of fine-tuning on HIV-1 protease (HIV-PR) protein/ligand complexes, with corresponding ground-truth fragments in the left most column. The fine-tuned model was trained to predict large fragments ( $\geq 4$  heavy atoms). The bottom five rows show the predictions of the model before finetuning when applied to the same examples. This foundational model was trained on large fragments (independent of chemical property or protein target). In both cases, the label set includes only large fragments, derived from the HIV-PR testing-set examples.

	Ground Truth	Top 5 Predictions				
Training: CA2 $\geq 4$ Label Set: CA2 $\geq 4$ Top-8: 47.6%		 1.000	 0.506	 0.455	 0.391	 0.389
		 0.615	 0.582	 0.566	 0.554	 0.553
		 1.000	 0.508	 0.455	 0.395	 0.390
		 0.551	 0.548	 0.543	 0.532	 0.524
		 0.805	 0.748	 0.697	 0.693	 0.653
Training: All $\geq 4$ , Epoch 60 Label Set: CA2 $\geq 4$ Top-8: 45.4%		 1.000	 0.506	 0.458	 0.393	 0.389
		 0.608	 0.586	 0.563	 0.530	 0.528
		 1.000	 0.506	 0.455	 0.391	 0.389
		 0.585	 0.556	 0.526	 0.517	 0.515
		 0.572	 0.560	 0.542	 0.521	 0.517

Figure S16. CA2 fine-tuning. The upper five rows show randomly selected testing-set predictions after 30 generations of fine-tuning on carbonic anhydrase 2 (CA2) protein/ligand complexes, with corresponding ground-truth fragments in the left most column. The fine-tuned model was trained to predict large fragments ( $\geq 4$  heavy atoms). The bottom five rows show the predictions of the model before finetuning when applied to the same examples. This foundational model was trained on large fragments (independent of chemical property or protein target). In both cases, the label set includes only large fragments, derived from the CA2 testing-set examples.

## References

- (1) Russell, R. J.; Haire, L. F.; Stevens, D. J.; Collins, P. J.; Lin, Y. P.; Blackburn, G. M.; Hay, A. J.; Gamblin, S. J.; Skehel, J. J. The structure of H5N1 avian influenza neuraminidase suggests new opportunities for drug design. *Nature* **2006**, *443* (7107), 45-49. DOI: 10.1038/nature05114 From NLM Medline.
- (2) Cheng, T. J.; Weinheimer, S.; Tarbet, E. B.; Jan, J. T.; Cheng, Y. S.; Shie, J. J.; Chen, C. L.; Chen, C. A.; Hsieh, W. C.; Huang, P. W.; et al. Development of oseltamivir phosphonate congeners as anti-influenza agents. *J Med Chem* **2012**, *55* (20), 8657-8670. DOI: 10.1021/jm3008486 From NLM Medline.
- (3) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J Chem Inf Model* **2010**, *50* (5), 742-754. DOI: 10.1021/ci100050t From NLM Medline.
- (4) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J Chem Inf Comput Sci* **2002**, *42* (6), 1273-1280. DOI: 10.1021/ci010132r From NLM Medline.