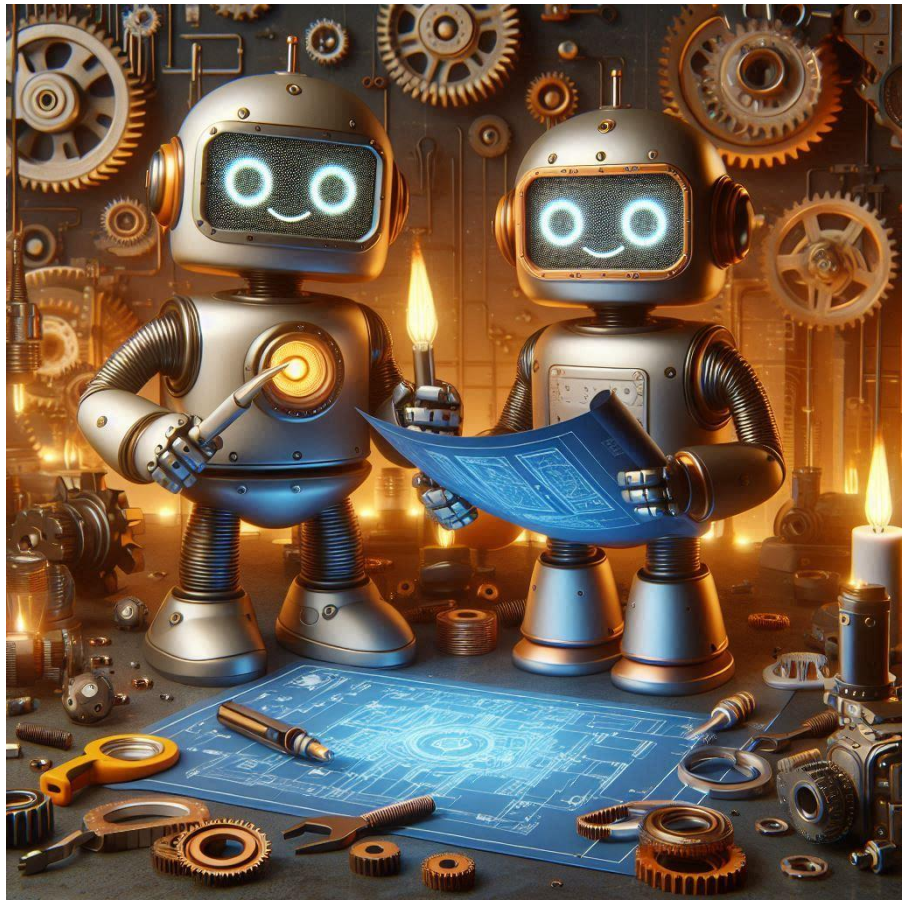
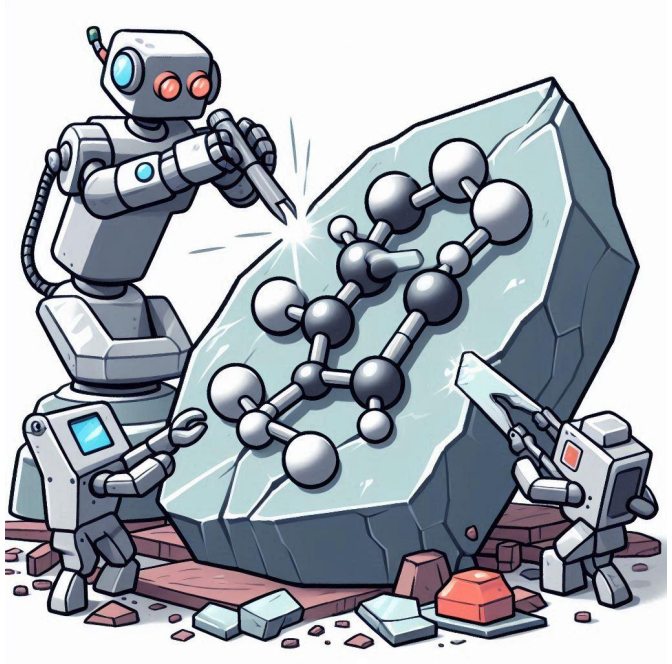


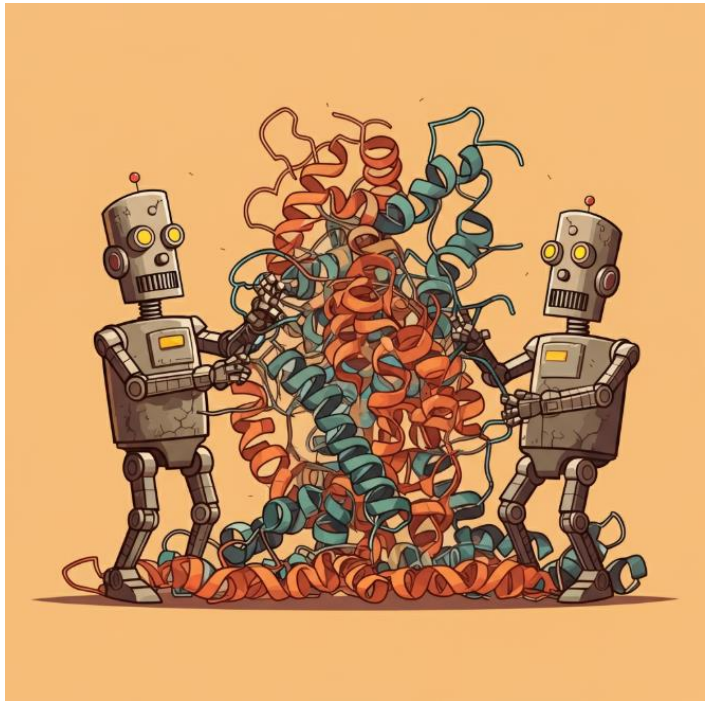
AI in Drug Discovery Workshop

Session 3 – Practical Applications

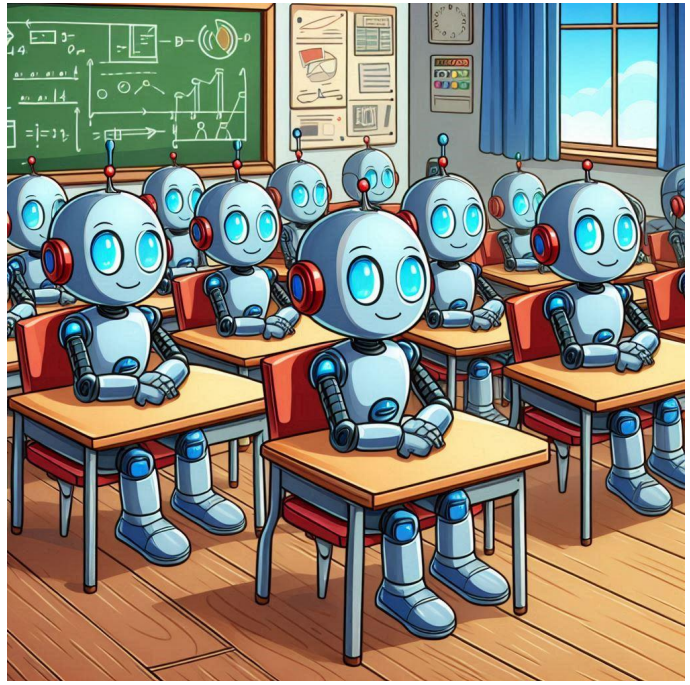
Practical Applications of Machine Learning in Drug Discovery



Molecule Generation

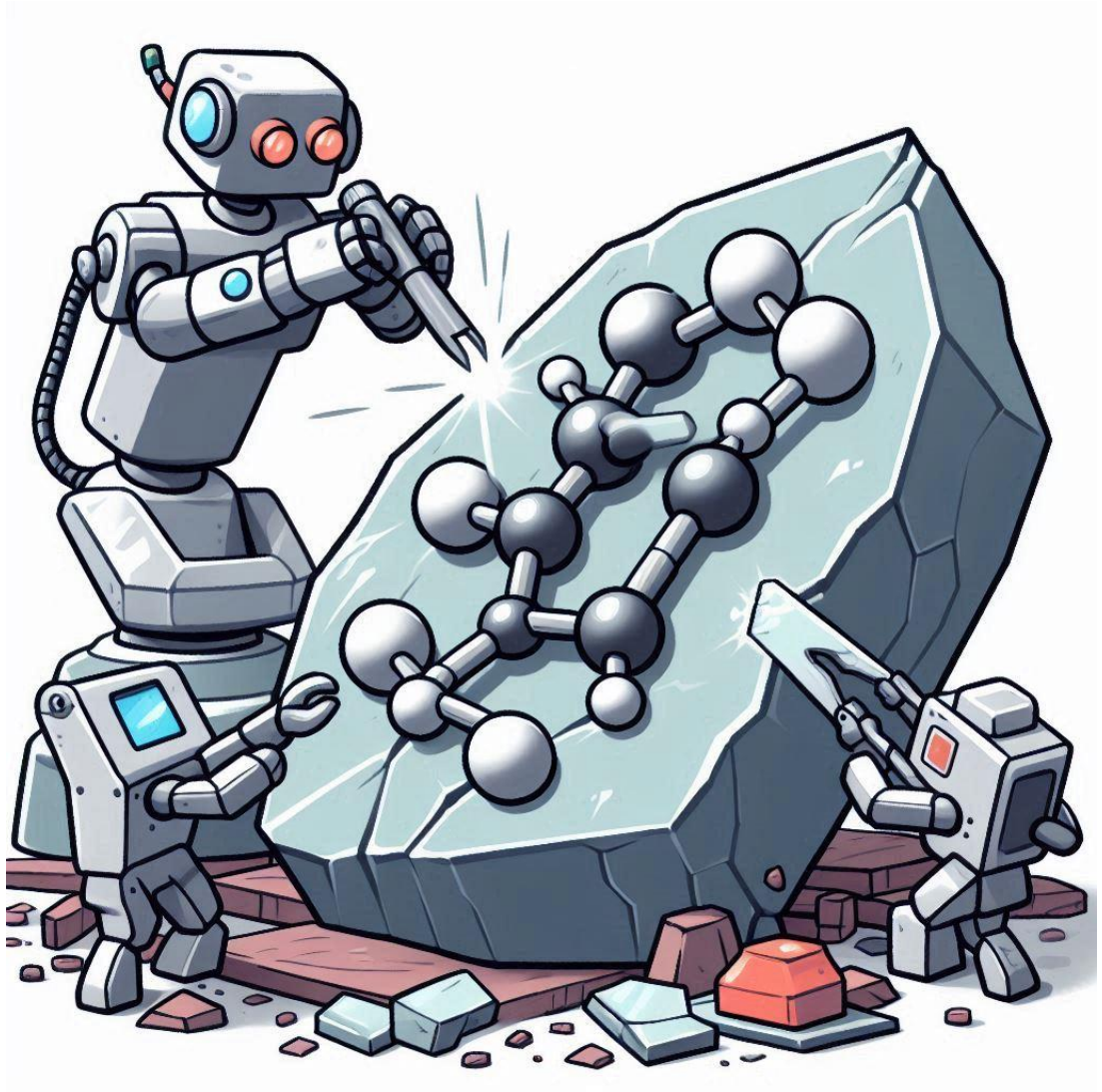


Protein Structure Prediction



Active Learning

Molecule Generation



Large Language Models (LLMs) Learn Next Token Probabilities

Given the phrase “The cat is”, what is the next token?

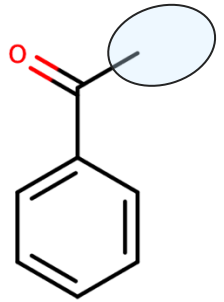
In this case, a token is a word

$$P(blue|The, cat, is) = .1$$
$$P(black|The, cat, is) = .85$$
$$P(green|The, cat, is) = .05$$

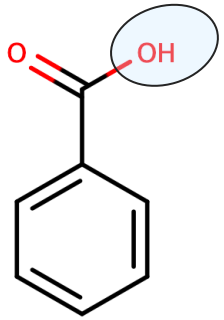
https://www.lakera.ai/blog/large-language-models-guide

Next Token Probabilities Can Also Be Applied to Chemistry

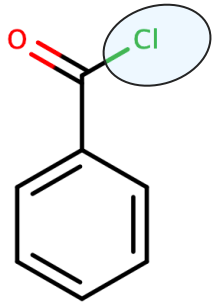
P(C,c1ccccc1C(=O)) = 0.2



P(O,c1ccccc1C(=O)) = 0.75



P(Cl,c1ccccc1C(=O)) = 0.05



Training a Recurrent Neural Network (RNN) to Generate “Shakespeare”

Prior

```
 First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famish?

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius is chief enemy to the people.

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own price.
Is't a verdict?

All:
No more talking on't; let it be done: away, away!

Second Citizen:
One word, good citizens.

First Citizen:
We are accounted poor
```

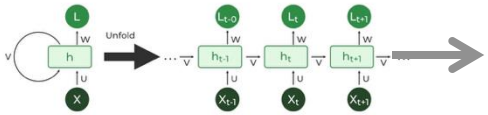
etc.

Embedding

```
array([18, 47, 56, 57, 58,  1, 15, 47, 58, 47, 64, 43, 52, 10,  0, 14, 43,
       44, 53, 56, 43,  1, 61, 43,  1, 54, 56, 53, 41, 43, 43, 42,  1, 39,
       52, 63,  1, 44, 59, 56, 58, 46, 43, 56,  6,  1, 46, 43, 39, 56,  1,
       51, 43,  1, 57, 54, 43, 39, 49,  8,  0,  0, 13, 50, 50, 10,  0, 31,
       54, 43, 39, 49,  6,  1, 57, 54, 43, 39, 49,  8,  0,  0, 18, 47, 56,
       57, 58,  1, 15, 47, 58, 47, 64, 43, 52, 10,  0, 37, 53, 59,  1, 39,
       56, 43,  1, 39, 50, 50,  1, 56, 43, 57, 53, 50, 60, 43, 42,  1, 56,
       39, 58, 46, 43, 56,  1, 58, 53,  1, 42, 47, 43,  1, 58, 46, 39, 52,
        1, 58, 53,  1, 44, 39, 51, 47, 57, 46, 12,  0,  0, 13, 50, 50, 10,
        0, 30, 43, 57, 53, 50, 60, 43, 42,  8,  1, 56, 43, 57, 53, 50, 60,
       43, 42,  8,  0,  0, 18, 47, 56, 57, 58,  1, 15, 47, 58, 47, 64, 43,
       52, 10,  0, 18, 47, 56, 57, 58,  6,  1, 63, 53, 59,  1, 49, 52, 53,
       61,  1, 15, 39, 47, 59, 57,  1, 25, 39, 56, 41, 47, 59, 57,  1, 47,
       57,  1, 41, 46, 47, 43, 44,  1, 43, 52, 43, 51, 63,  1, 58, 53,  1,
       58, 46, 43,  1, 54, 43, 53, 54, 50, 43,  8,  0,  0, 13, 50, 50, 10,
        0, 35, 43,  1, 49, 52, 53, 61,  5, 58,  6,  1, 61, 43,  1, 49, 52,
       53, 61,  5, 58,  8,  0,  0, 18, 47, 56, 57, 58,  1, 15, 47, 58, 47,
       64, 43, 52, 10,  0, 24, 43, 58,  1, 59, 57,  1, 49, 47, 50, 50,  1,
       46, 47, 51,  6,  1, 39, 52, 42,  1, 61, 43,  5, 50, 50,  1, 46, 39,
       60, 43,  1, 41, 53, 56, 52,  1, 39, 58,  1, 53, 59, 56,  1, 53, 61,
       52,  1, 54, 56, 47, 41, 43,  8,  0, 21, 57,  5, 58,  1, 39,  1, 60,
       43, 56, 42, 47, 41, 58, 12,  0,  0, 13, 50, 50, 10,  0, 26, 53,  1,
       51, 53, 56, 43,  1, 58, 39, 50, 49, 47, 52, 45,  1, 53, 52,  5, 58,
       11,  1, 50, 43, 58,  1, 47, 58,  1, 40, 43,  1, 42, 53, 52, 43, 10,
        1, 39, 61, 39, 63,  6,  1, 39, 61, 39, 63,  2,  0,  0, 31, 43, 41,
       53, 52, 42,  1, 15, 47, 58, 47, 64, 43, 52, 10,  0, 27, 52, 43,  1,
       61, 53, 56, 42,  6,  1, 45, 53, 53, 42,  1, 41, 47, 58, 47, 64, 43,
       52, 57,  8,  0,  0, 18, 47, 56, 57, 58,  1, 15, 47, 58, 47, 64, 43,
       52, 10,  0, 35, 43,  1, 39, 56, 43,  1, 39, 41, 41, 53, 59, 52, 58,
       43, 42,  1, 54, 53, 53, 56,  1, 41, 47, 58, 47, 64, 43, 52, 57,  6,
        1, 58, 46, 43,  1, 54, 39, 58, 56, 47, 41, 47, 39, 52, 57,  1, 45,
       53, 53, 42,  8,  0, 35, 46, 39, 58,  1, 39, 59, 58, 46, 53, 56, 47,
       58, 63,  1, 57, 59, 56, 44, 43, 47, 58, 57,  1, 53, 52,  1, 61, 53,
       59, 50, 42,  1, 56, 43, 50, 47, 43, 60, 43,  1, 59, 57, 10,  1, 47,
       44,  1, 58, 46, 43, 63,  0, 61, 53, 59, 50, 42,  1, 63, 47, 43, 50,
       42,  1, 59, 57,  1, 40, 59, 58,  1, 58, 46, 43,  1, 57, 59, 54, 43,
       56, 44, 50, 59, 47, 58, 63,  6,  1, 61, 46, 47, 50, 43,  1, 47, 58,
        1, 61, 43, 56, 43,  0, 61, 46, 53, 50, 43, 57, 53, 51, 43,  6,  1,
```

etc.

RNN



Output

```
ROMEO:
It mare to brave, my lord: good peace! Feeping to be crown'd,
I dreamt occapity: a baudish, make thee mellow.

VINCENTIO:
I' the duke, sir, their beaming on a despite of more
view to the prozont; and my vow,
Tuguse, when dreadful Norts of Hills,
Unless you be the moonest malice or tent to false;
whistle and hours are too much commended.

JULIET:
Dost thou not, but have they may come there.

ESCALUS:
We have: I'll visit with thee
And bid Bimine own fight and beggarly; whom
Too little liking, for I know thou art not vice,
Though father is no less a man that left your high
As twenty thousand end,
Wrangening so well as eat him by changing
Against the Duke of York, Summon are me:
I am she, poor soul, I envy nothing feast,
Alack the heaven gails of mine heir toes
was in the view of Juliet o' the seas;
And every gash of invition, but not
His hands received the estremest pass
And safely with my hands.
```

etc.

Training a Recurrent Neural Network (RNN) to Generate “Shakespeare”

Prior

```
 First Citizen:
Before we proceed any further, hear me speak.

All:
Speak, speak.

First Citizen:
You are all resolved rather to die than to famis

All:
Resolved. resolved.

First Citizen:
First, you know Caius Marcius is chief enemy to

All:
We know't, we know't.

First Citizen:
Let us kill him, and we'll have corn at our own
Is't a verdict?

All:
No more talking on't; let it be done: away, away

Second Citizen:
One word, good citizens.

First Citizen:
We are accounted poor
```

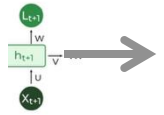
etc.

```
ROMEO:
It mare to brave, my lord: good peace! Feeping to be crown'd,
I dreamt occapity: a baudish, make thee mellow.

VINCENTIO:
I' the duke, sir, their beaming on a despite of more
view to the prozont; and my vow,
Tuguse, when dreadful Norts of Hills,
Unless you be the moonest malice or tent to false;
whistle and hours are too much commended.

JULIET:
Dost thou not, but have they may come there.

ESCALUS:
We have: I'll visit with thee
And bid Bimine own fight and beggarly; whom
Too little liking, for I know thou art not vice,
Though father is no less a man that left your high
As twenty thousand end,
Wrangening so well as eat him by changing
Against the Duke of York, Summon are me:
I am she, poor soul, I envy nothing feast,
Alack the heaven gails of mine heir toes
was in the view of Juliet o' the seas;
And every gash of invition, but not
His hands received the estremest pass
And safely with my hands.
```



Output

```
ROMEO:
It mare to brave, my lord: good peace! Feeping to be crown'd,
I dreamt occapity: a baudish, make thee mellow.

VINCENTIO:
I' the duke, sir, their beaming on a despite of more
view to the prozont; and my vow,
Tuguse, when dreadful Norts of Hills,
Unless you be the moonest malice or tent to false;
whistle and hours are too much commended.

JULIET:
Dost thou not, but have they may come there.

ESCALUS:
We have: I'll visit with thee
And bid Bimine own fight and beggarly; whom
Too little liking, for I know thou art not vice,
Though father is no less a man that left your high
As twenty thousand end,
Wrangening so well as eat him by changing
Against the Duke of York, Summon are me:
I am she, poor soul, I envy nothing feast,
Alack the heaven gails of mine heir toes
was in the view of Juliet o' the seas;
And every gash of invition, but not
His hands received the estremest pass
And safely with my hands.
```

etc.

Training a Recurrent Neural Network (RNN) to Generate SMILES

Prior

```
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1Cl
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(C#N)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)cc(C)c1C(O)c1ccc(Cl)cc1
Cc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1
Cc1cc(Br)ccc1C(=O)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C(c1ccc(Cl)cc1Cl)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
CS(=O)(=O)c1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)cc1
c1cc2cc(c1)-c1cccc(c1)C[n+]1ccc(c3ccccc31)NCCCCCCCCCCNc1cc[n+](c3ccccc13)C2
CSc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)cc1
O=C(c1ccc(Cl)cc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C1O/C(=C/Br)CCC1c1ccccc2ccccc12
O=C1O/C(=C/I)CCC1c1ccccc2ccccc12
C#C/C=C1\CCC(c2cccc3ccccc23)C(=O)O1
O=C1O/C(=C\I)CCC1c1ccccc2ccccc12
O=C(c1ccccc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1
O=c1cnn(-c2ccc(C(O)c3ccc(Cl)cc3)c(Cl)c2)c(=O)[nH]1
O=C(c1ccccc1)c1cc(-n2ncc(=O)[nH]c2=O)ccc1Cl
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(O)c1ccccc1
COc1cc2nc(N3CCN(C(=O)C4CC4c4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
COc1cc2nc(N3CCN(S(=O)(=O)c4cccs4)CC3)nc(N)c2cc1OC
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)-c2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2cccc(-c3cccc(C[n+]4ccc(N)c5ccccc54)c3)c2)c2ccccc12
Nc1cc[n+](Cc2ccc(CCc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
Nc1cc[n+](Cc2ccc(/C=C\c3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
c1ccc(C[n+]2ccc(NCCCCCCCCCCNc3cc[n+](Cc4ccccc4)c4ccccc34)c3ccccc32)cc1
C1=C\c2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCNc2cc[n+](c3ccccc23)Cc2ccc/1cc2
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)CCc2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)Cc2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2ccc3c(c2)Cc2cc(C[n+]4ccc(N)c5ccccc54)ccc2-3)c2ccccc12
Nc1cc[n+](Cc2ccc(-c3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1Cl
c1ccc2c(c1)c1cc[n+]2Cc2ccc3c(c2)Cc2cc(ccc2-3)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2ccc(Cc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
COC(=O)/C=C/C(=O)Nc1ccc(CCCCC(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(C(=O)CCc4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(C(=O)/C=C/c4cccc(F)c4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)C4CC5C=CC4C5)CC3)nc(N)c2cc1OC
C/C=C/C=C/C(=O)N1CCN(c2nc(N)c3cc(OC)c(OC)cc3n2)CC1
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccccc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(F)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(Cl)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C(F)=C/c4ccccc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)Nc4ccccc4)CC3)nc(N)c2cc1OC
COC(=O)/C=C/C(=O)Nc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(/C(S)=N/C4CCCCC4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)C4=CC5C=CC4CC5)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(S(=O)(=O)c4ccc(C)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(/C(S)=N/c4ccccc4)CC3)nc(N)c2cc1OC
```

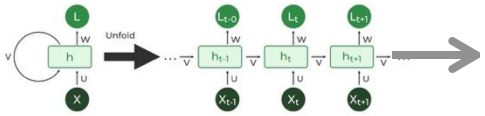
etc.

Embedding

```
array([25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41,
       51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 46, 12, 25,  5, 21, 34,  6,
       46, 12, 46, 46, 46, 46, 46, 46, 12, 25, 50,  2, 25, 28, 26, 32, 24, 31,
       17, 14, 13, 20,  1, 25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,
        5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 12,
       25,  5, 21, 34,  6, 46, 12, 46, 46, 46,  5, 25,  3, 33,  6, 46, 46,
       12,  2, 25, 28, 26, 32, 24, 31, 17, 14, 13, 19,  1, 25, 46, 12, 46,
       46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46,
       13, 21, 34,  6, 46, 46,  5, 25,  6, 46, 12, 25,  5, 34,  6, 46, 12,
       46, 46, 46,  5, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32, 24, 31,
       13, 17, 16, 17, 17, 18,  1, 25, 46, 12, 46, 46, 46,  5, 25,  5, 21,
       34,  6, 46, 13, 46, 46, 46,  5,  8, 51, 14, 51, 46, 46,  5, 21, 34,
        6, 41, 51, 28, 43, 46, 14, 21, 34,  6, 46, 46, 13,  6, 46, 46, 12,
        2, 25, 28, 26, 32, 24, 31, 17, 14, 17, 13,  1, 25, 46, 12, 46, 46,
        5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13,
       21, 34,  6, 46, 46, 46, 12, 25,  5, 21, 34,  6, 46, 12, 46, 46, 46,
        5, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32, 24, 31, 13, 17, 18,
       19, 17, 15,  1, 25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5,
       21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 12, 25,
        5, 21, 34,  6, 46, 12, 46, 46, 46, 46, 46, 12,  2, 25, 28, 26, 32,
       24, 31, 17, 14, 17, 14,  1, 25, 46, 12, 46, 46,  5, 24, 54,  6, 46,
       46, 46, 12, 25,  5, 21, 34,  6, 46, 12, 46, 46, 46,  5,  8, 51, 13,
       51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46,
       46, 12, 25, 50,  2, 25, 28, 26, 32, 24, 31, 17, 14, 16, 13,  1, 34,
       21, 25,  5, 46, 12, 46, 46, 46,  5, 25, 50,  6, 46, 46, 12, 25, 50,
        6, 46, 12, 46, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6,
       41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 12, 25, 50,  2, 25, 28,
       26, 32, 24, 31, 13, 17, 19, 11, 20, 18,  1, 25, 37,  5, 21, 34,  6,
        5, 21, 34,  6, 46, 12, 46, 46, 46, 46,  5, 25,  5, 21, 34,  6, 46, 13,
       46, 46, 46,  5,  8, 51, 14, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28,
       43, 46, 14, 21, 34,  6, 46, 46, 13, 25, 50,  6, 46, 46, 12,  2, 25,
       28, 26, 32, 24, 31, 13, 17, 17, 15, 16, 18,  1, 46, 12, 46, 46, 13,
       46, 46,  5, 46, 12,  6,  8, 46, 12, 46, 46, 46,  5, 46, 12,  6,
       25, 41, 51,  7, 43, 12, 46, 46, 46,  5, 46, 14, 46, 46, 46, 46, 46,
       14, 12,  6, 33, 25, 25, 25, 25, 25, 25, 25, 25, 33, 46, 12,
       46, 46, 41, 51,  7, 43,  5, 46, 14, 46, 46, 46, 46, 12, 14,  6,
       25, 13,  2, 25, 28, 26, 32, 24, 31, 13, 17, 19, 16, 16, 17,  1, 25,
       37, 46, 12, 46, 46, 46,  5, 25,  5, 21, 34,  6, 46, 13, 46, 46, 46,
        5,  8, 51, 14, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 14,
       21, 34,  6, 46, 46, 13, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32,
       24, 31, 17, 13, 12, 15,  1, 34, 21, 25,  5, 46, 12, 46, 46, 46,  5,
       25, 50,  6, 46, 46, 12,  6, 46, 12, 46, 46, 46,  5,  8, 51, 13, 51,
       46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46,
       12, 25, 50,  2, 25, 28, 26, 32, 24, 31, 17, 13, 13, 13,  1, 34, 21,
```

etc.

RNN



Output

```
CC(=O)Oc1c(-c2ccc(F)cc2)ncn1Cc1ccc(F)cc1
CC(C)CC(NCCc1ccc(O)cc1)c1ccc(Cl)cc1Cl
CCOc1ccc(NC(=O)N2c3ccccc3CC2C)cc1OCC
O=C(NC1(c2ccccc2)CCCC1)c1cc(-c2sccc2Cl)ccc1Br
OCC1OC(n2cc(F)c3c(SCc4ccc(Cl)c(Cl)c4)ncnc32)C(O)C1O
N#CC(=Cc1cccc(OCC(=O)O)c1)c1nc2ccccc2[nH]1
CC(=O)OCC1(OC)CCC2(C)C3=CC=C4C(CCC(OC50C(CO)C(O)C(O)C5O)C(O)C40)C3(C)CCC23CC13CCC(=O)O
CCCN(CCC)C1CCc2c(C)cccc2Cl
CNS(=O)(=O)Cc1nsc(NCc2ccccc2)n1
CCOC(=O)Cl=C(C)NC2=C(C(=O)CC(c3ccc(OC)cc3)C2)C1c1ccccc(F)c1
CC1C(=O)N(c2nccc(C#Cc3cccc(F)c3)c2)C(=O)N1CCO
O=Cc1cc2c([nH]c3ccccc32)c(-c2cccc(O)c2)n1
Cc1ccc(C2NCC(O)C2CO)cc1
COc1cc2nc(NCCc3ccccn3)nc(N)c2cc1OC
Brc1ccc2[nH]cc(C3CCC3)c2c1
O=C(NC1C2CC3CC1CC(O)(C3)C2)c1ccc(NC2CCS(=O)(=O)C2)nn1Cc1ccccc1
OC1CCCN(Cc2ccc(C(=O)NCCN3CCCC3)cc2)C1
CCOc1ccc(OCCCC(=O)C(C)C)c(OC)c1
CC(=O)NC(CSC(=O)N(CCC#N)CCCCN)C(=O)O
COc1ccccc1-c1nnc(C(C)Nc2ccc(CCNCC(O)c3cccc(N)c3)cc2)o1
CCSc1nc(Nc2ccc(C#N)c(C(F)(F)F)c2)nc(N2CCOCC2)n1
CS(=O)(=O)N1CCC(CNC(=O)c2cc(Cl)c(Cl)s2)CC1
COC(=O)C=CC(CC(=O)N(CCC1)N=O)NCCOCCOCCOCCOCCNCCCCCCCNCCCCCCCCCNCCCN
COC(=O)C(NC(=O)C(=O)Nc1ccccc1OC)C1OC(C(=O)OCC=Cc2ccc(-c3ccccc3)cc2)C(C)(C)S1
CN1C(=O)C(NC(=O)Nc2ccc(Cl)cc2)N=C(c2ccc(Br)cc2)c2ccccc21
CCCCN(CCCC)CCOc1ccc(Cl)cc1Cl
CCCS(=O)(=O)N(CN)c1nc2cc(OC)ccc2s1
COc1cc(-c2ccc3c(c2)NC(=O)C3=Cc2[nH]c(C)c(C)c2C)ccc1O
CC1=C(CCl)C(=O)CC2C=CN(C)C(=O)C12
CCOC(=O)c1cc(COc2ccc3nc(C(=O)OC)oc3c2)c(=O)[nH]1
Cc1c(CNC2CCCCCC2)nc(-c2ccc(Br)cc2)n1-c1ccc(Cl)cc1
CCCC(CO)Oc1ccc(-c2cc(C(F)(F)F)nc(NC(CC)CO)n2)cc1
CCN(CC)c1ccc(NC(=O)Cc2coc3cc(OC)ccc23)cc1
OCC1OC(n2c(=O)[nH]cc(O)c2=S)CCO1
CCC(O)(CCSCCCC)COP(=O)(O)OP(=O)(O)O
CCN(CC)S(=O)(=O)c1cccc(NC(=O)c2ccc(C)c(S(=O)(=O)N3CCOCC3)c2)c1
COc1ccccc1OCc1cc(C(=O)NC2CCCCC2)nn1-c1ccccc1
OCC1OC(n2c(NC3CC3)nc3sc4c(OCCF)nncc4c3nc2=O)CC1O
Cc1c(-c2cccnc2)nc(C(C)C)n2c1nc1ccc(C)cc12
O=C(C1CCOCC1)N1CCN(c2ncnc(Cl)c2Br)CC1
CC(=O)c1cc(C(C)(C)C)sc1NC(=O)C(C)(C)c1ccc(C(C)(C)C)c(=O)[nH]n1
CN=C(S)N1CCC(c2cccc(C(F)(F)F)c2)CC1
COc1ccc(NC(=O)c2ccccc2Cl)cc1S(=O)(=O)N1CCOCC1
c1ccc(C#Cc2c[nH]c3ncnc(N4CCOCC4)c23)cc1
COc1ccc(CN2CCOC3CN(C(C)=O)CCC32)cc1OC
CC1(C)OCC(NCC(=O)Nc2ccc(Sc3ccc(Cl)cc3C(=O)N3CCOCC3)cc2)C1
```

etc.

Training a Recurrent Neural Network (RNN) to Generate SMILES

Prior

```
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1Cl
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(C#N)cc1
Cclcc(-n2ncc(=O)[nH]c2=O)cc(C)c1C(O)c1ccc(Cl)cc1
Cclccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1
Cclcc(Br)ccc1C(=O)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C(c1ccc(Cl)cc1Cl)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
CS(=O)(=O)c1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)
c1cc2cc(c1)-c1cccc(c1)C[n+]1ccc(c3ccccc31)NCCCCCCCCC
CSc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)cc1
O=C(c1ccc(Cl)cc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C1O/C(=C/Br)CCC1c1cccc2ccccc12
O=C1O/C(=C/I)CCC1c1cccc2ccccc12
C#C/C=C1\CCC(c2cccc3ccccc23)C(=O)O1
O=C1O/C(=C\I)CCC1c1cccc2ccccc12
O=C(c1ccccc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1
O=c1cnn(-c2ccc(C(O)c3ccc(Cl)cc3)c(Cl)c2)c(=O)[nH]1
O=C(c1ccccc1)c1cc(-n2ncc(=O)[nH]c2=O)ccc1Cl
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(O)c1ccccc1
COc1cc2nc(N3CCN(C(=O)C4CC4c4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
COc1cc2nc(N3CCN(S(=O)(=O)c4cccs4)CC3)nc(N)c2cc1OC
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)-c2ccc(cc2)C[n+]2ccc(
Nc1cc[n+](Cc2cccc(-c3cccc(C[n+]4ccc(N)c5ccccc54)c3)c
Nc1cc[n+](Cc2ccc(CCc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)c
Nc1cc[n+](Cc2ccc(/C=C\c3ccc(C[n+]4ccc(N)c5ccccc54)cc
c1ccc(C[n+]2ccc(NCCCCCCCCCCNc3cc[n+](C4ccccc4)c4ccc
C1=C\c2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCNc2cc(
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)CCc2ccc(cc2)C[n+]2ccc
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)Cc2ccc(cc2)C[n+]2ccc(
Nc1cc[n+](Cc2ccc3c(c2)Cc2cc(C[n+]4ccc(N)c5ccccc54)cc
Nc1cc[n+](Cc2ccc(-c3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc
Cclcc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1Cl
c1ccc2c(c1)c1cc[n+]2Cc2ccc3c(c2)Cc2cc(ccc2-3)C[n+]2c
Nc1cc[n+](Cc2ccc(Cc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc
COC(=O)/C=C/C(=O)Nc1ccc(CCCCC(=O)N2CCN(c3nc(N)c4cc(O
COc1cc2nc(N3CCN(C(=O)CCc4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC
COc1cc2nc(N3CCN(C(=O)/C=C/c4cccc(F)c4)CC3)nc(N)c2cc1
COc1cc2nc(N3CCN(C(=O)C4CC5C=CC4C5)CC3)nc(N)c2cc1OC
C/C=C/C=C/C(=O)N1CCN(c2nc(N)c3cc(OC)c(OC)cc3n2)CC1
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccccc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(F)cc4)CC3)nc(N)c2cc1
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(Cl)cc4)CC3)nc(N)c2cc
COc1cc2nc(N3CCN(C(=O)/C(F)=C/c4ccccc4)CC3)nc(N)c2cc1
COc1cc2nc(N3CCN(C(=O)Nc4ccccc4)CC3)nc(N)c2cc1OC
COC(=O)/C=C/C(=O)Nc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(
COc1cc2nc(N3CCN(/C(S)=N/C4CCCCC4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)C4=CC5C=CC4CC5)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(S(=O)(=O)c4ccc(C)cc4)CC3)nc(N)c2cc1O
COc1cc2nc(N3CCN(/C(S)=N/c4ccccc4)CC3)nc(N)c2cc1OC
```

etc.

Output

```
CC(=O)Oc1c(-c2ccc(F)cc2)ncn1Cc1ccc(F)cc1
CC(C)CC(NCCc1ccc(O)cc1)c1ccc(Cl)cc1Cl
CCOc1ccc(NC(=O)N2c3ccccc3CC2C)cc1OCC
O=C(NC1(c2ccccc2)CCCC1)c1cc(-c2sccc2Cl)ccc1Br
OCC1OC(n2cc(F)c3c(SCc4ccc(Cl)c(Cl)c4)ncnc32)C(O)C1O
N#CC(=Cc1cccc(OCC(=O)O)c1)c1nc2ccccc2[nH]1
CC(=O)OCC1(OC)CCC2(C)C3=CC=C4C(CCC(OC50C(CO)C(O)C(O)C50)C(O)C40)C3(C)CCC23CC13CCC(=O)O
CCCN(CCC)C1CCc2c(C)cccc2Cl
CNS(=O)(=O)Cc1nsc(NCc2ccccc2)n1
CCOC(=O)Cl=C(C)NC2=C(C(=O)CC(c3ccc(OC)cc3)C2)C1c1ccccc(F)c1
CC1C(=O)N(c2nccc(C#Cc3cccc(F)c3)c2)C(=O)N1CCO
O=Cc1cc2c([nH]c3ccccc32)c(-c2cccc(O)c2)n1
Cc1ccc(C2NCC(O)C2CO)cc1
COc1cc2nc(NCCc3ccccn3)nc(N)c2cc1OC
Brc1ccc2[nH]cc(C3CCC3)c2c1
O=C(NC1C2CC3CC1CC(O)(C3)C2)c1cc(NC2CCS(=O)(=O)C2)nn1Cc1ccccc1
OC1CCCN(Cc2ccc(C(=O)NCCN3CCCC3)cc2)C1
CCOc1ccc(OCCCC(=O)C(C)C)c(OC)c1
CC(=O)NC(CSC(=O)N(CCC#N)CCCCN)C(=O)O
COc1ccccc1-c1nnc(C(C)Nc2ccc(CCNCC(O)c3cccc(N)c3)cc2)o1
CCSc1nc(Nc2ccc(C#N)c(C(F)(F)F)c2)nc(N2CCOCC2)n1
CS(=O)(=O)N1CCC(CNC(=O)c2cc(Cl)c(Cl)s2)CC1
COC(=O)C=CC(CC(=O)N(CCC1)N=O)NCCOCCOCCOCCOCCNCCCCCCCNCCCCCCCCCNCCCN
COC(=O)C(NC(=O)C(=O)Nc1ccccc1OC)C1OC(C(=O)OCC=Cc2ccc(-c3ccccc3)cc2)C(C)(C)S1
CN1C(=O)C(NC(=O)Nc2ccc(Cl)cc2)N=C(c2ccc(Br)cc2)c2ccccc21
CCCCN(CCCC)CCOc1ccc(Cl)cc1Cl
CCCS(=O)(=O)N(CN)c1nc2cc(OC)ccc2s1
COc1cc(-c2ccc3c(c2)NC(=O)C3=Cc2[nH]c(C)c(C)c2C)ccc1O
CC1=C(CCl)C(=O)CC2C=CN(C)C(=O)C12
CCOC(=O)c1cc(COc2ccc3nc(C(=O)OC)oc3c2)c(=O)[nH]1
Cc1c(CNC2CCCCCC2)nc(-c2ccc(Br)cc2)n1-c1ccc(Cl)cc1
CCCC(CO)Oc1ccc(-c2cc(C(F)(F)F)nc(NC(CC)CO)n2)cc1
CCN(CC)c1ccc(NC(=O)Cc2coc3cc(OC)ccc23)cc1
OCC1OC(n2c(=O)[nH]cc(O)c2=S)CCO1
CCC(O)(CCSCCCC)COP(=O)(O)OP(=O)(O)O
CCN(CC)S(=O)(=O)c1cccc(NC(=O)c2ccc(C)c(S(=O)(=O)N3CCOCC3)c2)c1
COc1ccccc1OCc1cc(C(=O)NC2CCCCC2)nn1-c1ccccc1
OCC1OC(n2c(NC3CC3)nc3sc4c(OCCF)nncc4c3nc2=O)CC1O
Cc1c(-c2cccnc2)nc(C(C)C)n2c1nc1ccc(C)cc12
O=C(C1CCOCC1)N1CCN(c2ncnc(Cl)c2Br)CC1
CC(=O)c1cc(C(C)(C)C)sc1NC(=O)C(C)(C)c1ccc(C(C)(C)C)c(=O)[nH]n1
CN=C(S)N1CCC(c2cccc(C(F)(F)F)c2)CC1
COc1ccc(NC(=O)c2ccccc2Cl)cc1S(=O)(=O)N1CCOCC1
c1ccc(C#Cc2c[nH]c3ncnc(N4CCOCC4)c23)cc1
COc1ccc(CN2COC3CN(C(C)=O)CCC32)cc1OC
CC1(C)OCC(NCC(=O)Nc2ccc(Sc3ccc(Cl)cc3C(=O)N3CCOCC3)cc2)C1
```

etc.

Augmented Likelihood Incorporates Scoring Into Molecule Generation

Prior

```
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1Cl
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(C#N)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)cc(C)c1C(O)c1ccc(Cl)cc1
Cc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccccc1
Cc1cc(Br)ccc1C(=O)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C(c1ccc(Cl)cc1Cl)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
CS(=O)(=O)c1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)cc1
c1cc2cc(c1)-c1cccc(c1)C[n+]1ccc(c3ccccc31)NCCCCCCCCCCNc1cc[n+](c3ccccc13)C2
CSc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2Cl)cc1
O=C(c1ccc(Cl)cc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1Cl
O=C1O/C(=C/Br)CCC1c1ccccc2ccccc12
O=C1O/C(=C/I)CCC1c1ccccc2ccccc12
C#C/C=C1\CCC(c2cccc3ccccc23)C(=O)O1
O=C1O/C(=C\I)CCC1c1ccccc2ccccc12
O=C(c1ccccc1)c1ccc(-n2ncc(=O)[nH]c2=O)cc1
O=c1cnn(-c2ccc(C(O)c3ccc(Cl)cc3)c(Cl)c2)c(=O)[nH]1
O=C(c1ccccc1)c1cc(-n2ncc(=O)[nH]c2=O)ccc1Cl
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(O)c1ccccc1
COc1cc2nc(N3CCN(C(=O)C4CC4c4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(C(=O)c2ccc(-n3ncc(=O)[nH]c3=O)cc2)cc1
COc1cc2nc(N3CCN(S(=O)(=O)c4cccs4)CC3)nc(N)c2cc1OC
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)-c2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2ccc(-c3cccc(C[n+]4ccc(N)c5ccccc54)c3)c2)c2ccccc12
Nc1cc[n+](Cc2ccc(CCc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
Nc1cc[n+](Cc2ccc(/C=C\c3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
c1ccc(C[n+]2ccc(NCCCCCCCCCCNc3cc[n+](Cc4ccccc4)c4ccccc34)c3ccccc32)cc1
C1=C\c2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCNc2cc[n+](c3ccccc23)Cc2ccc/1cc2
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)CCc2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
c1ccc2c(c1)c1cc[n+]2Cc2ccc(cc2)Cc2ccc(cc2)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2ccc3c(c2)Cc2cc(C[n+]4ccc(N)c5ccccc54)ccc2-3)c2ccccc12
Nc1cc[n+](Cc2ccc(-c3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
Cc1cc(-n2ncc(=O)[nH]c2=O)ccc1C(=O)c1ccc(Cl)cc1Cl
c1ccc2c(c1)c1cc[n+]2Cc2ccc3c(c2)Cc2cc(ccc2-3)C[n+]2ccc(c3ccccc32)NCCCCCCCCCCN1
Nc1cc[n+](Cc2ccc(Cc3ccc(C[n+]4ccc(N)c5ccccc54)cc3)cc2)c2ccccc12
COC(=O)/C=C/C(=O)Nc1ccc(CCCCC(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(C(=O)CCc4ccccc4)CC3)nc(N)c2cc1OC
COc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(C(=O)/C=C/c4cccc(F)c4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)C4CC5C=CC4C5)CC3)nc(N)c2cc1OC
C/C=C/C=C/C(=O)N1CCN(c2nc(N)c3cc(OC)c(OC)cc3n2)CC1
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccccc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(F)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C=C/c4ccc(Cl)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)/C(F)=C/c4ccccc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)Nc4ccccc4)CC3)nc(N)c2cc1OC
COC(=O)/C=C/C(=O)Nc1ccc(/C=C/C(=O)N2CCN(c3nc(N)c4cc(OC)c(OC)cc4n3)CC2)cc1
COc1cc2nc(N3CCN(/C(S)=N/C4CCCCC4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(C(=O)C4=CC5C=CC4CC5)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(S(=O)(=O)c4ccc(C)cc4)CC3)nc(N)c2cc1OC
COc1cc2nc(N3CCN(/C(S)=N/c4ccccc4)CC3)nc(N)c2cc1OC
```

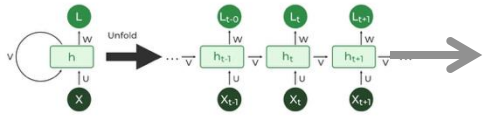
etc.

Embedding

```
array([25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41,
       51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 12, 25,  5, 21, 34,  6,
       46, 12, 46, 46, 46, 46, 46, 12, 25, 50,  2, 25, 28, 26, 32, 24, 31,
       17, 14, 13, 20,  1, 25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,
        5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 12,
       25,  5, 21, 34,  6, 46, 12, 46, 46, 46,  5, 25,  3, 33,  6, 46, 46,
       12,  2, 25, 28, 26, 32, 24, 31, 17, 14, 13, 19,  1, 25, 46, 12, 46,
       46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46,
       13, 21, 34,  6, 46, 46,  5, 25,  6, 46, 12, 25,  5, 34,  6, 46, 12,
       46, 46, 46,  5, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32, 24, 31,
       13, 17, 16, 17, 17, 18,  1, 25, 46, 12, 46, 46, 46,  5, 25,  5, 21,
       34,  6, 46, 13, 46, 46, 46,  5,  8, 51, 14, 51, 46, 46,  5, 21, 34,
        6, 41, 51, 28, 43, 46, 14, 21, 34,  6, 46, 46, 13,  6, 46, 46, 12,
        2, 25, 28, 26, 32, 24, 31, 17, 14, 17, 13,  1, 25, 46, 12, 46, 46,
        5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13,
       21, 34,  6, 46, 46, 46, 12, 25,  5, 21, 34,  6, 46, 12, 46, 46, 46,
        5, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32, 24, 31, 13, 17, 18,
       19, 17, 15,  1, 25, 46, 12, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5,
       21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 46, 12, 25,
        5, 21, 34,  6, 46, 12, 46, 46, 46, 46, 46, 12,  2, 25, 28, 26, 32,
       24, 31, 17, 14, 17, 14,  1, 25, 46, 12, 46, 46,  5, 24, 54,  6, 46,
       46, 46, 12, 25,  5, 21, 34,  6, 46, 12, 46, 46, 46,  5,  8, 51, 13,
       51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46,
       46, 12, 25, 50,  2, 25, 28, 26, 32, 24, 31, 17, 14, 16, 13,  1, 34,
       21, 25,  5, 46, 12, 46, 46, 46,  5, 25, 50,  6, 46, 46, 12, 25, 50,
        6, 46, 12, 46, 46, 46,  5,  8, 51, 13, 51, 46, 46,  5, 21, 34,  6,
       41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46, 12, 25, 50,  2, 25, 28,
       26, 32, 24, 31, 13, 17, 19, 11, 20, 18,  1, 25, 37,  5, 21, 34,  6,
        5, 21, 34,  6, 46, 12, 46, 46, 46,  5, 25,  5, 21, 34,  6, 46, 13,
       46, 46, 46,  5,  8, 51, 14, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28,
       43, 46, 14, 21, 34,  6, 46, 46, 13, 25, 50,  6, 46, 46, 12,  2, 25,
       28, 26, 32, 24, 31, 13, 17, 17, 15, 16, 18,  1, 46, 12, 46, 46, 13,
       46, 46,  5, 46, 12,  6,  8, 46, 12, 46, 46, 46,  5, 46, 12,  6,
       25, 41, 51,  7, 43, 12, 46, 46, 46,  5, 46, 14, 46, 46, 46, 46, 46,
       14, 12,  6, 33, 25, 25, 25, 25, 25, 25, 25, 25, 25, 33, 46, 12,
       46, 46, 41, 51,  7, 43,  5, 46, 14, 46, 46, 46, 46, 46, 12, 14,  6,
       25, 13,  2, 25, 28, 26, 32, 24, 31, 13, 17, 19, 16, 16, 17,  1, 25,
       37, 46, 12, 46, 46, 46,  5, 25,  5, 21, 34,  6, 46, 13, 46, 46, 46,
        5,  8, 51, 14, 51, 46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 14,
       21, 34,  6, 46, 46, 13, 25, 50,  6, 46, 46, 12,  2, 25, 28, 26, 32,
       24, 31, 17, 13, 12, 15,  1, 34, 21, 25,  5, 46, 12, 46, 46, 46,  5,
       25, 50,  6, 46, 46, 12,  6, 46, 12, 46, 46, 46,  5,  8, 51, 13, 51,
       46, 46,  5, 21, 34,  6, 41, 51, 28, 43, 46, 13, 21, 34,  6, 46, 46,
       12, 25, 50,  2, 25, 28, 26, 32, 24, 31, 17, 13, 13, 13,  1, 34, 21,
```

etc.

RNN



Output

```
CC(=O)Oc1c(-c2ccc(F)cc2)ncn1Cc1ccc(F)cc1
CC(C)CC(NCCc1ccc(O)cc1)c1ccc(Cl)cc1Cl
CCOc1ccc(NC(=O)N2c3ccccc3CC2C)cc1OCC
O=C(NC1(c2ccccc2)CCCC1)c1cc(-c2sccc2Cl)ccc1Br
OCC1OC(n2cc(F)c3c(SCc4ccc(Cl)c(Cl)c4)ncnc32)C(O)C1O
N#CC(=Cc1cccc(OCC(=O)O)c1)c1nc2ccccc2[nH]1
CC(=O)OCC1(OC)CCC2(C)C3=CC=C4C(CCC(OC50C(CO)C(O)C(O)C50)C(O)C40)C3(C)CCC23CC13CCC(=O)O
CCCN(CCC)C1CCc2c(C)cccc2Cl
CNS(=O)(=O)Cc1nsc(NCc2ccccc2)n1
CCOC(=O)Cl=C(C)NC2=C(C(=O)CC(c3ccc(OC)cc3)C2)C1c1ccccc(F)c1
CC1C(=O)N(c2ncnc(C#Cc3cccc(F)c3)c2)C(=O)N1CCO
O=Cc1cc2c([nH]c3ccccc32)c(-c2cccc(O)c2)n1
Cc1ccc(C2NCC(O)C2CO)cc1
COc1cc2nc(NCCc3ccccn3)nc(N)c2cc1OC
Brc1ccc2[nH]cc(C3CCC3)c2c1
O=C(NC1C2CC3CC1CC(O)(C3)C2)c1cc(NC2CCS(=O)(=O)C2)nn1Cc1ccccc1
OC1CCCN(Cc2ccc(C(=O)NCCN3CCCC3)cc2)C1
CCOc1ccc(OCCCC(=O)C(C)C)c(OC)c1
CC(=O)NC(CSC(=O)N(CCC#N)CCCCN)C(=O)O
COc1ccccc1-c1nnc(C(C)Nc2ccc(CCNCC(O)c3cccc(N)c3)cc2)o1
CCSc1nc(Nc2ccc(C#N)c(C(F)(F)F)c2)nc(N2CCOCC2)n1
CS(=O)(=O)N1CCC(CNC(=O)c2cc(Cl)c(Cl)s2)CC1
COC(=O)C=CC(CC(=O)N(CCC1)N=O)NCCOCCOCCOCCOCCNCCCCCCNCCCCCCCCCNCCCN
COC(=O)C(NC(=O)C(=O)Nc1ccccc1OC)C1OC(C(=O)OCC=Cc2ccc(-c3ccccc3)cc2)C(C)(C)S1
CN1C(=O)C(NC(=O)Nc2ccc(Cl)cc2)N=C(c2ccc(Br)cc2)c2ccccc21
CCCCN(CCCC)CCOc1ccc(Cl)cc1Cl
CCCS(=O)(=O)N(CN)c1nc2cc(OC)ccc2s1
COc1cc(-c2ccc3c(c2)NC(=O)C3=Cc2[nH]c(C)c(C)c2C)ccc1O
CC1=C(CCl)C(=O)CC2C=CN(C)C(=O)C12
CCOC(=O)c1cc(COc2ccc3nc(C(=O)OC)oc3c2)c(=O)[nH]1
Cc1c(CNC2CCCCCC2)nc(-c2ccc(Br)cc2)n1-c1ccc(Cl)cc1
CCCC(CO)Oc1ccc(-c2cc(C(F)(F)F)nc(NC(CC)CO)n2)cc1
CCN(CC)c1ccc(NC(=O)Cc2coc3cc(OC)ccc23)cc1
OCC1OC(n2c(=O)[nH]cc(O)c2=S)CCO1
CCC(O)(CCSCCCC)COP(=O)(O)OP(=O)(O)O
CCN(CC)S(=O)(=O)c1cccc(NC(=O)c2ccc(C)c(S(=O)(=O)N3CCOCC3)c2)c1
COc1ccccc1OCc1cc(C(=O)NC2CCCCC2)nn1-c1ccccc1
OCC1OC(n2c(NC3CC3)nc3sc4c(OCCF)nncc4c3nc2=O)CC1O
Cc1c(-c2cccnc2)nc(C(C)C)n2c1nc1ccc(C)cc12
O=C(C1CCOCC1)N1CCN(c2ncnc(Cl)c2Br)CC1
CC(=O)c1cc(C(C)(C)C)sc1NC(=O)C(C)(C)c1ccc(C(C)(C)C)c(=O)[nH]n1
CN=C(S)N1CCC(c2cccc(C(F)(F)F)c2)CC1
COc1ccc(NC(=O)c2ccccc2Cl)cc1S(=O)(=O)N1CCOCC1
c1ccc(C#Cc2c[nH]c3ncnc(N4CCOCC4)c23)cc1
COc1ccc(CN2CCOC3CN(C(C)=O)CCC32)cc1OC
CC1(C)OCC(NCC(=O)Nc2ccc(Sc3ccc(Cl)cc3C(=O)N3CCOCC3)cc2)C1
```

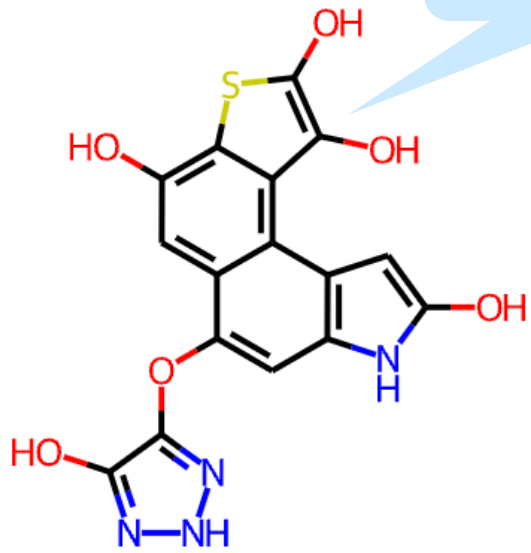
etc.

Augmented likelihood

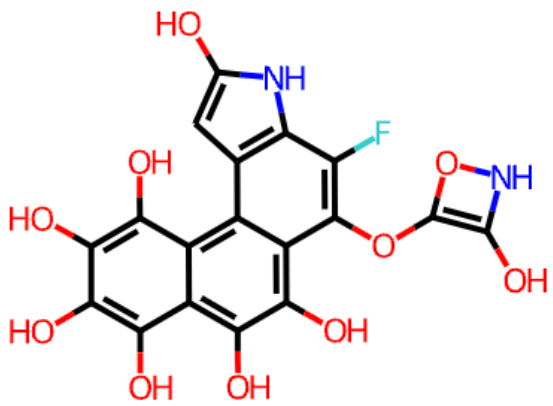
$$log\ P(A)_U = log\ P(A)_{Prior} + \sigma S(A)$$

Generative Molecules Can “Game” Scoring Functions

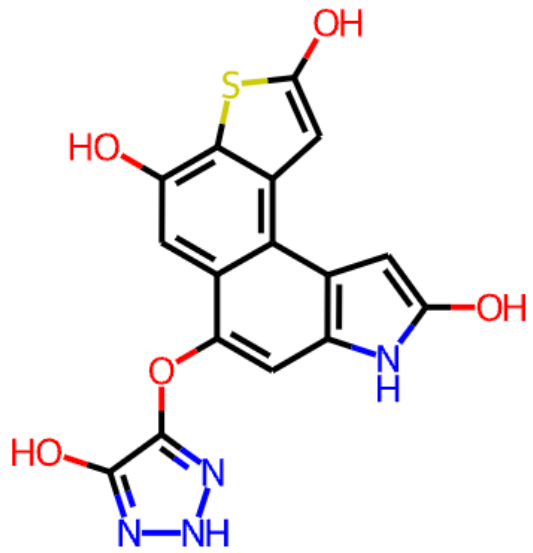
Unprecedented in ChEMBL



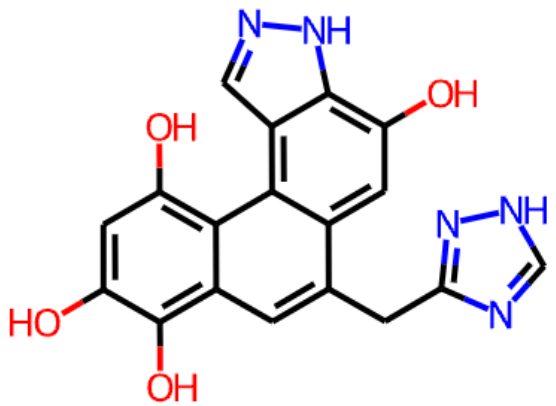
21.96



21.60



21.43

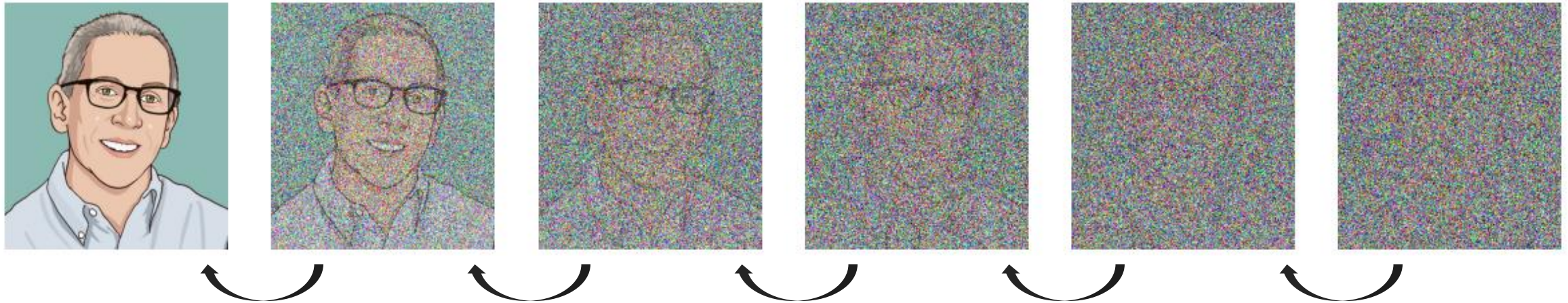


21.40

Dyachenko, Natalia V., et al. "Synthesis of fused heterocyclic systems via the Mallory photoreaction of arylthienylethenes." *Photochemical & Photobiological Sciences* 18.12 (2019): 2901-2911.

Diffusion Models Learn to Successively Denoise Data

Successive addition of Gaussian noise



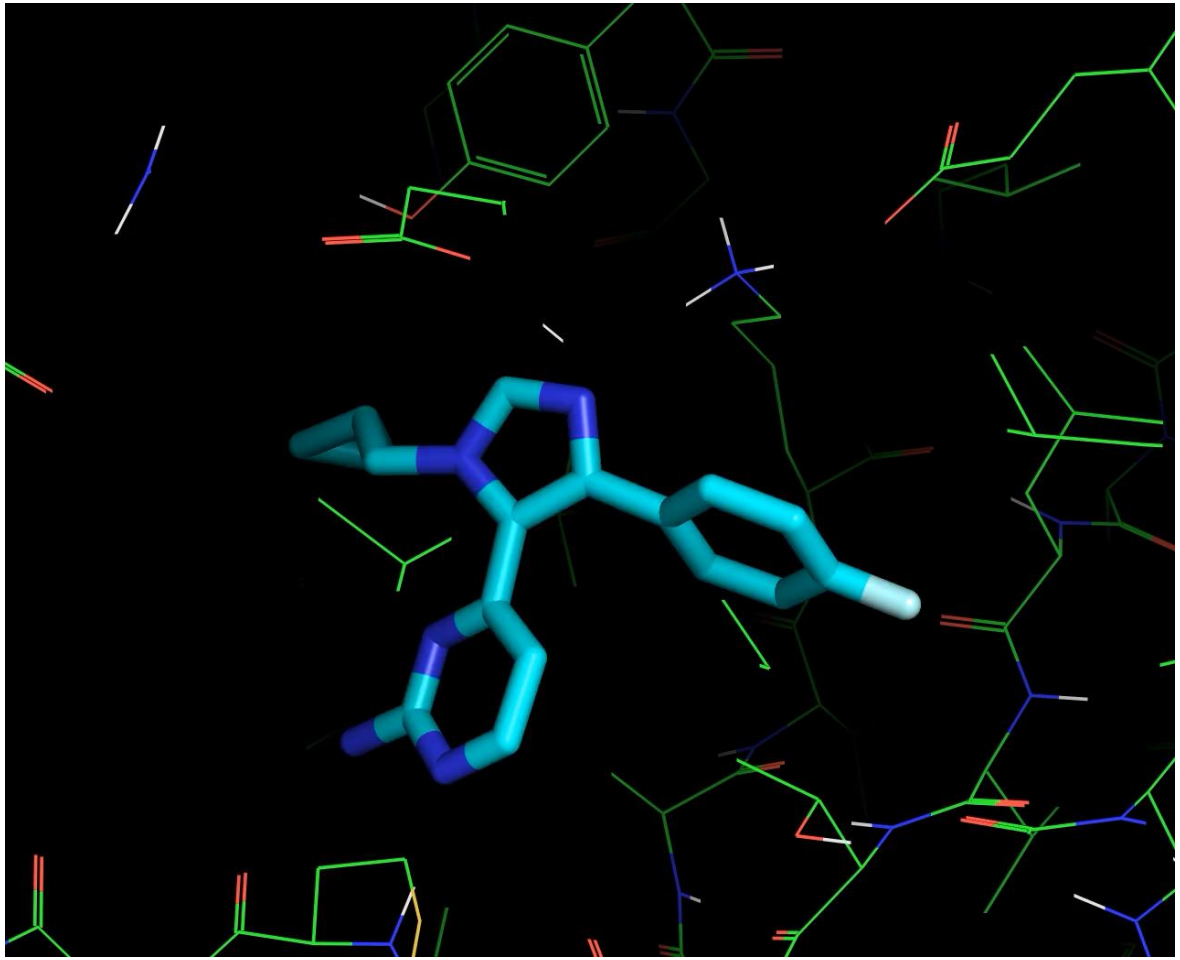
Data

Noise

Learn to reverse the process

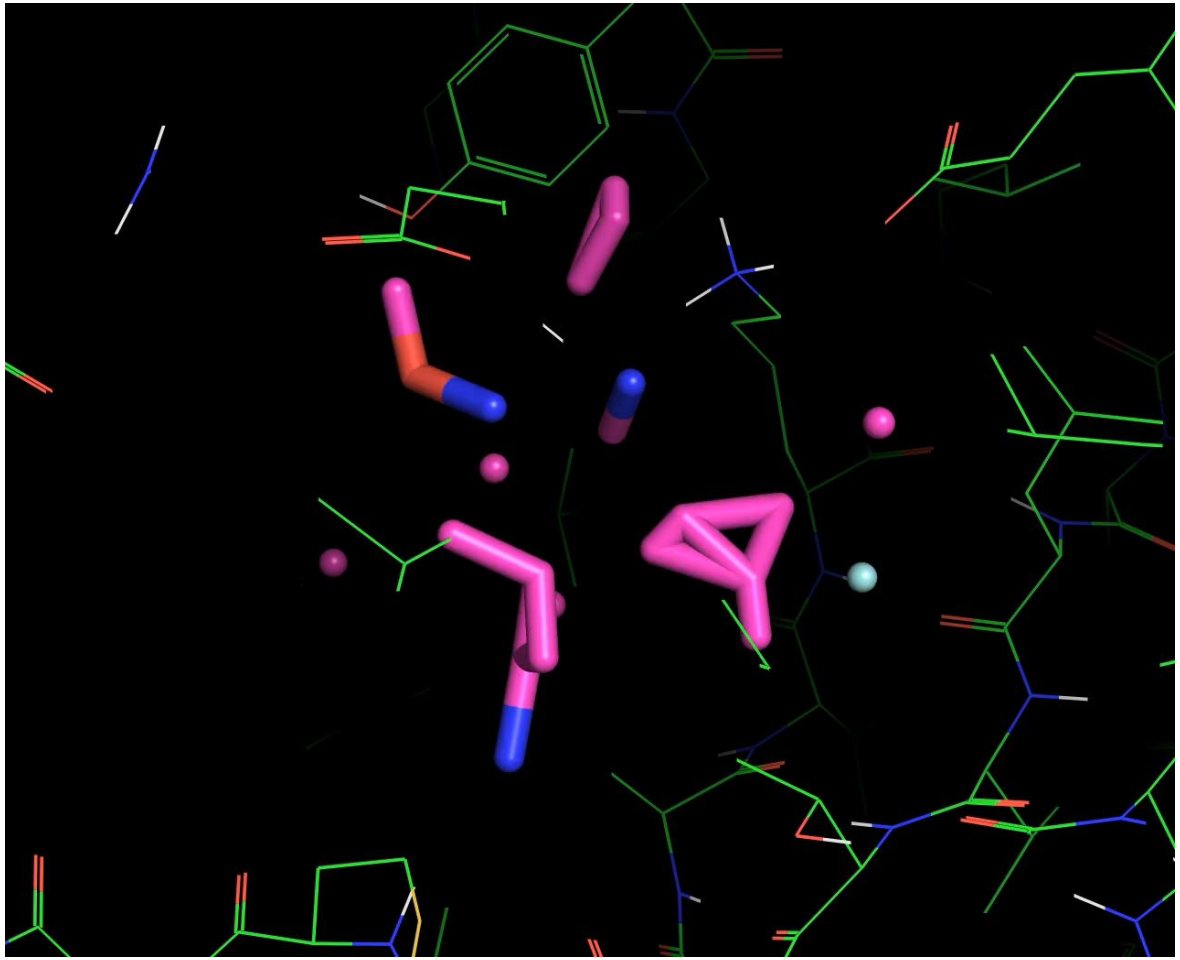
Once trained, the model can generate “data” from noise

Diffusion Can Also Be Used To Generate 3D Protein and Ligand Structures



Similar denoising strategies can be applied to atoms in a binding site

A Trained Diffusion Model Can Transform Atoms at Random Positions



Begin with atoms at random positions and use diffusion to generate a bound molecule

Molecule Generation is Not Without Its Problems

molecules

MDPI

Article

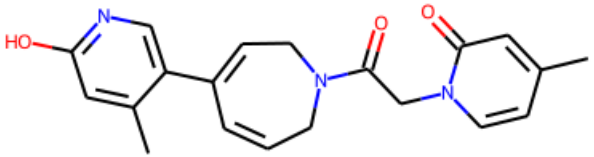
Design of SARS-CoV-2 Main Protease Inhibitors Using Artificial Intelligence and Molecular Dynamic Simulations

Lars Elend [1], Luise Jacobsen [2], Tim Cofala [1], Jonas Prellberg [1], Thomas Teusch [3], Oliver Kramer [1,*] and Ilia A. Solov'yov [3,4,5,*]

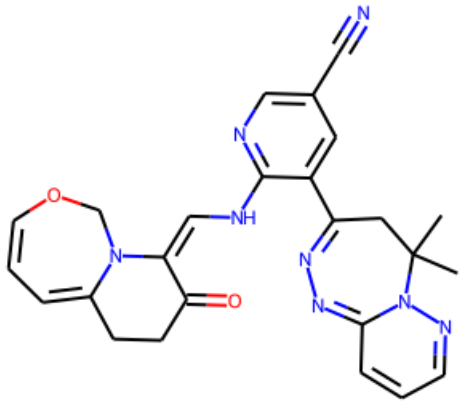
1 Computational Intelligence Lab, Department of Computer Science, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; lars.elend@uni-oldenburg.de (L.E.); tim.cofala@uni-oldenburg.de (T.C.); jonas.prellberg@uni-oldenburg.de (J.P.)
2 Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; luja@sdu.dk
3 Department of Physics, Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany; thomas.teusch@uni-oldenburg.de
4 Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany
5 Center for Nanoscale Dynamics (CENAD), Carl von Ossietzky Universität Oldenburg, Institut für Physik, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
* Correspondence: oliver.kramer@uol.de (O.K.); ilia.solovyov@uni-oldenburg.de (I.A.S.); Tel.: +49-441-798-3817 (I.A.S.)

Molecules 2022, **27**, 4020

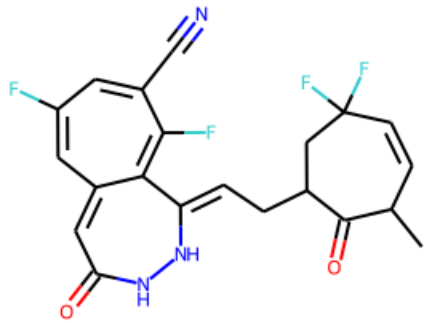
Highest Scoring Molecules Contain Questionable Functionality



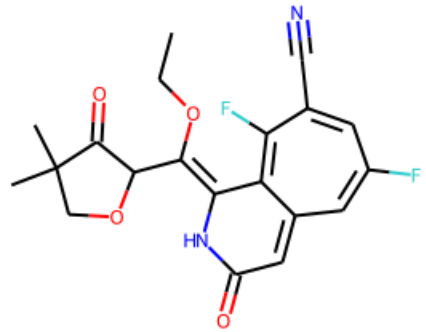
Lig 1



Lig 7

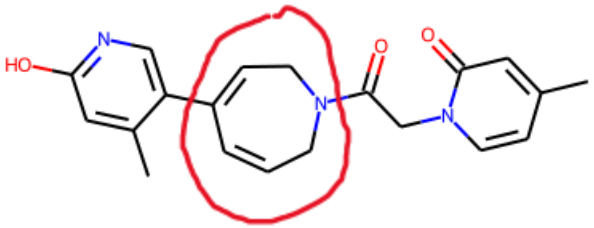


Lig 19

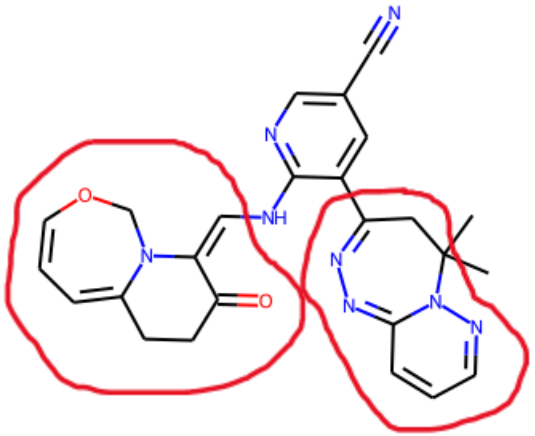


Lig 21

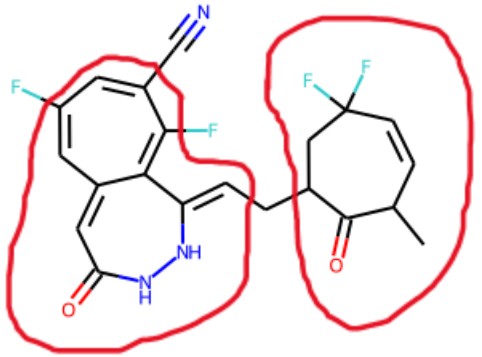
Highest Scoring Molecules Contain Questionable Functionality



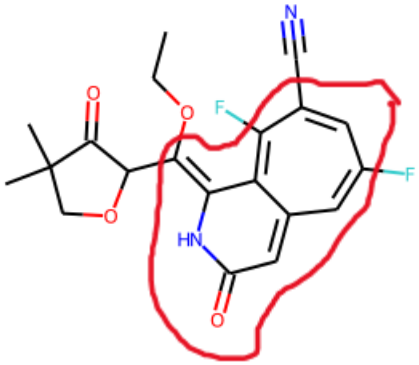
Lig 1



Lig 7

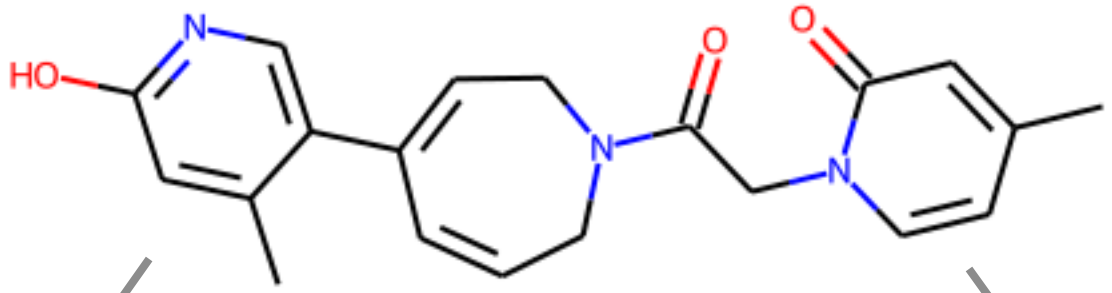


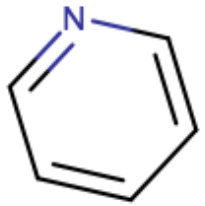
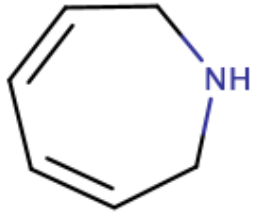
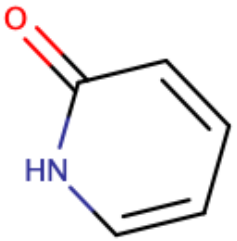
Lig 19



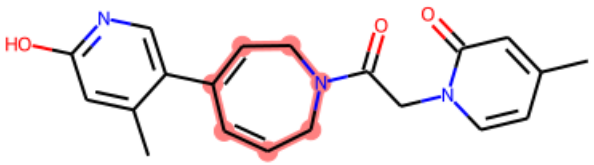
Lig 21

Evaluating Ring System Frequency

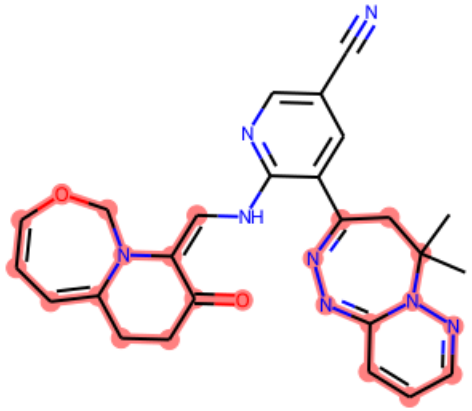


Ring			
Occurrences in ChEMBL	225,475	0	11,363
	✅	❌	✅

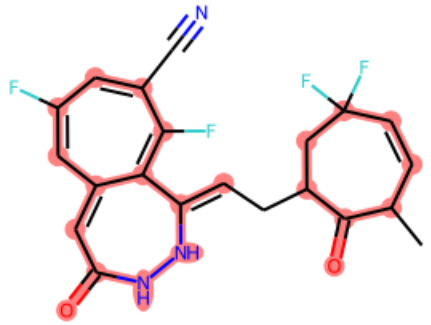
Top Molecules



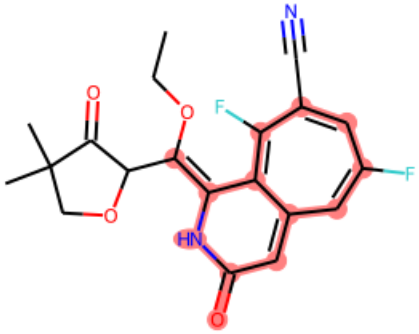
Lig 1



Lig 7



Lig 19



Lig 21

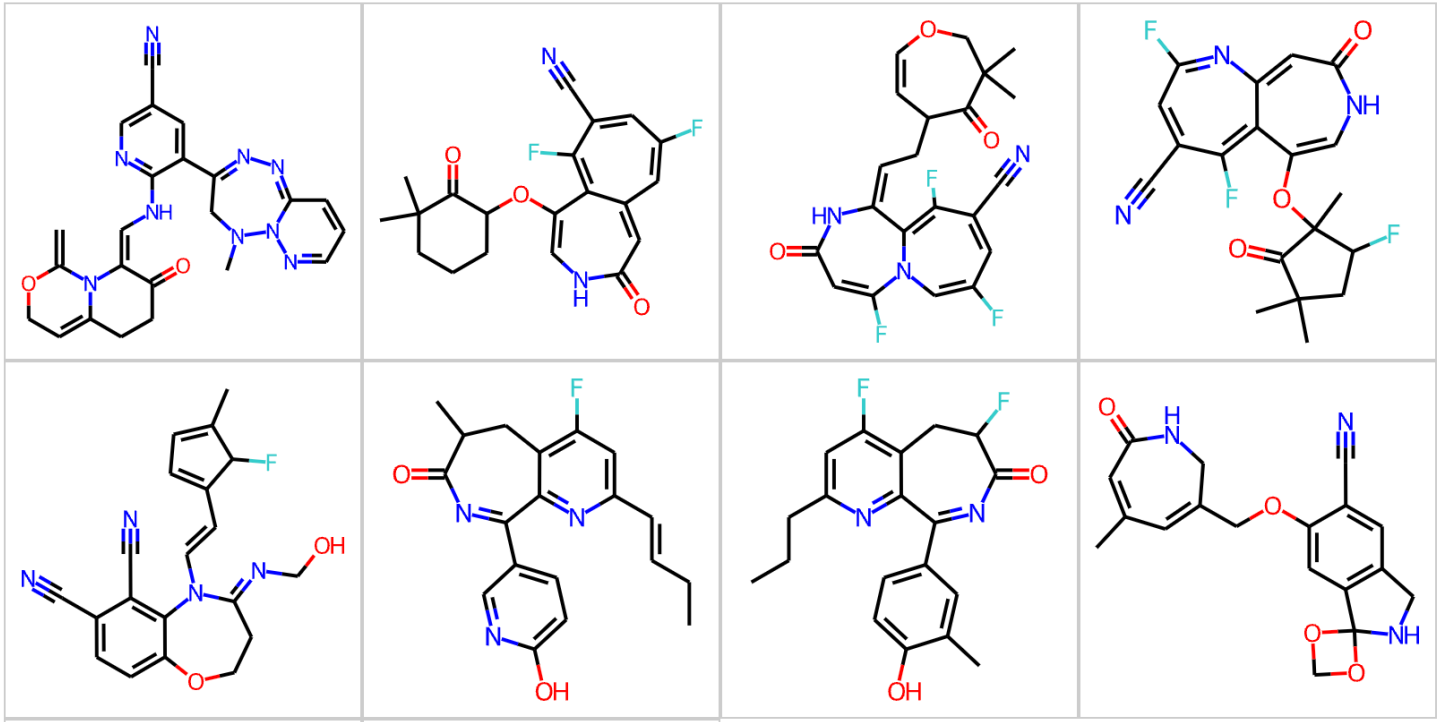
Evaluating Generative Model Output

144,350 molecules generated

23 molecules violated rules of valence

107,386 (74%) contained ring systems not found in ChEMBL

144,350 (79%) contained ring systems occurring < 10 times in ChEMBL



molecules

MDPI

Article

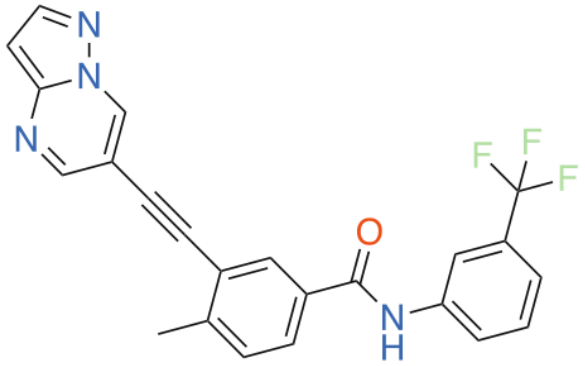
Design of SARS-CoV-2 Main Protease Inhibitors Using Artificial Intelligence and Molecular Dynamic Simulations

Lars Elend [1], Luise Jacobsen [2], Tim Cofala [1], Jonas Prellberg [1], Thomas Teusch [3], Oliver Kramer [1,*] and Ilia A. Solov'yov [3,4,5,*]

1 Computational Intelligence Lab, Department of Computer Science, Carl von Ossietzky University, Ammerländer Heerstraße 114-118, 26129 Oldenburg, Germany; lars.elend@uni-oldenburg.de (L.E.); tim.cofala@uni-oldenburg.de (T.C.); jonas.prellberg@uni-oldenburg.de (J.P.)
2 Department of Physics, Chemistry and Pharmacy, University of Southern Denmark, Campusvej 55, 5230 Odense M, Denmark; luja@sdu.dk
3 Department of Physics, Carl von Ossietzky University, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany; thomas.teusch@uni-oldenburg.de
4 Research Center for Neurosensory Science, Carl von Ossietzky Universität Oldenburg, 26111 Oldenburg, Germany
5 Center for Nanoscale Dynamics (CENAD), Carl von Ossietzky Universität Oldenburg, Institut für Physik, Ammerländer Heerstr. 114-118, 26129 Oldenburg, Germany
* Correspondence: oliver.kramer@uol.de (O.K.); ilia.solovyov@uni-oldenburg.de (I.A.S.); Tel.: +49-441-798-3817 (I.A.S.)

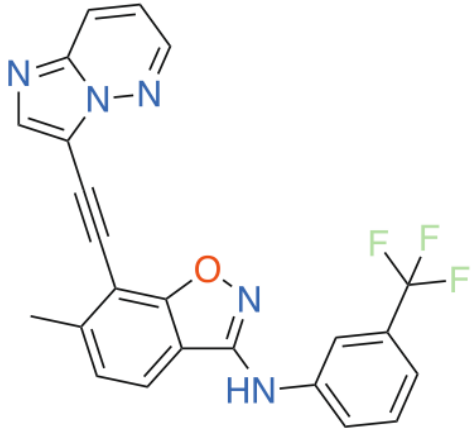
Molecules 2022, **27**, 4020

Assessing the Novelty of Molecules Produced By Generative Models



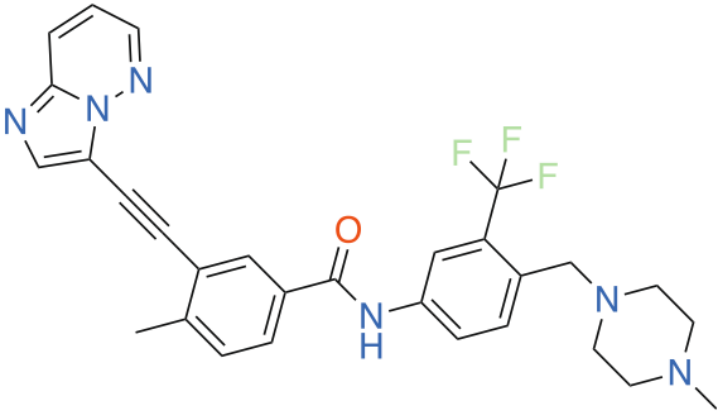
Gao et al.
Compound 7r
6 nM

Published in JMC 2013



Zhavoronkov et al.
Compound 1
10 nM

Generative Design
2020

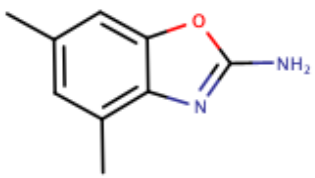


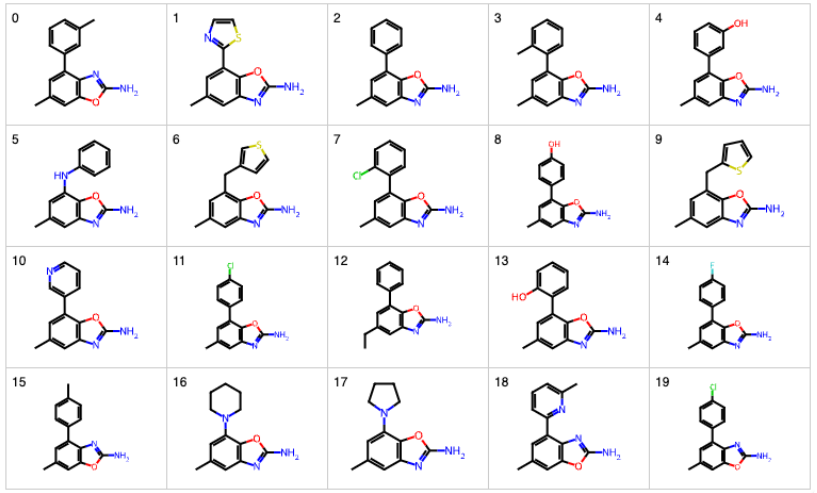
Ponatinib
9 nM

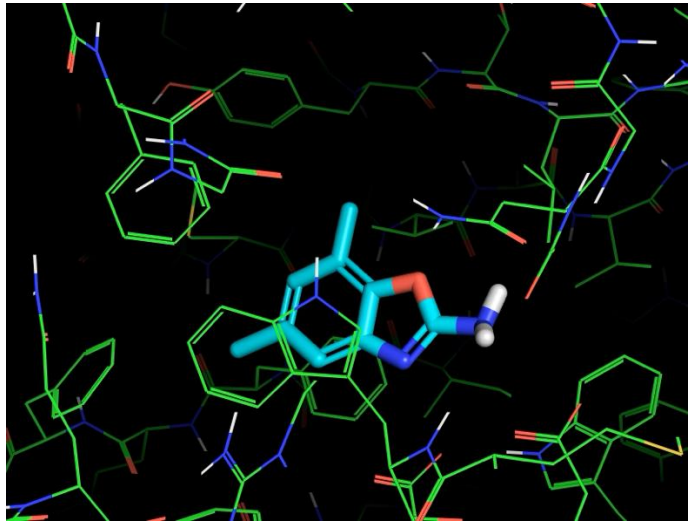
FDA Approved
2012

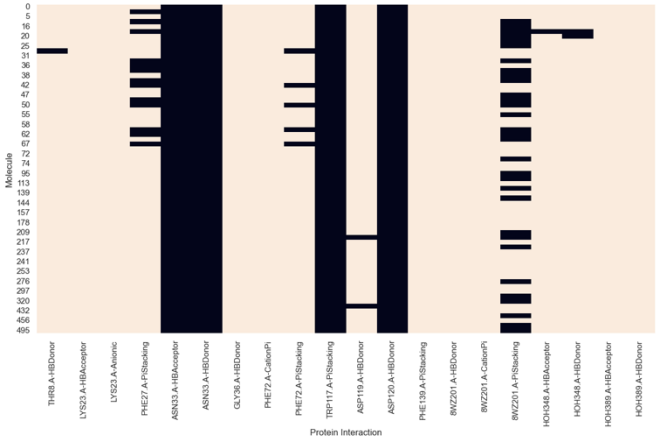
Walters, Murcko, *Nature Biotechnology*, 2020

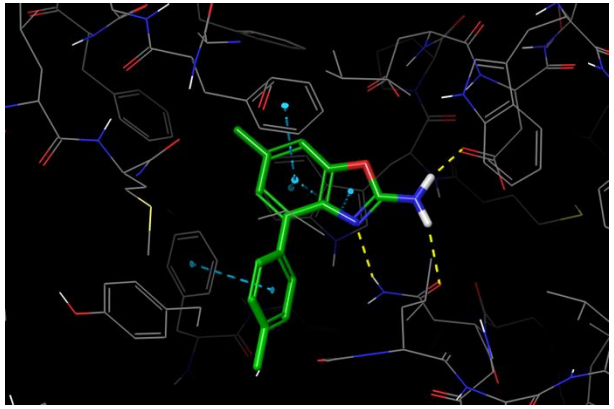
A Generative Design Workflow for Fragment Screening



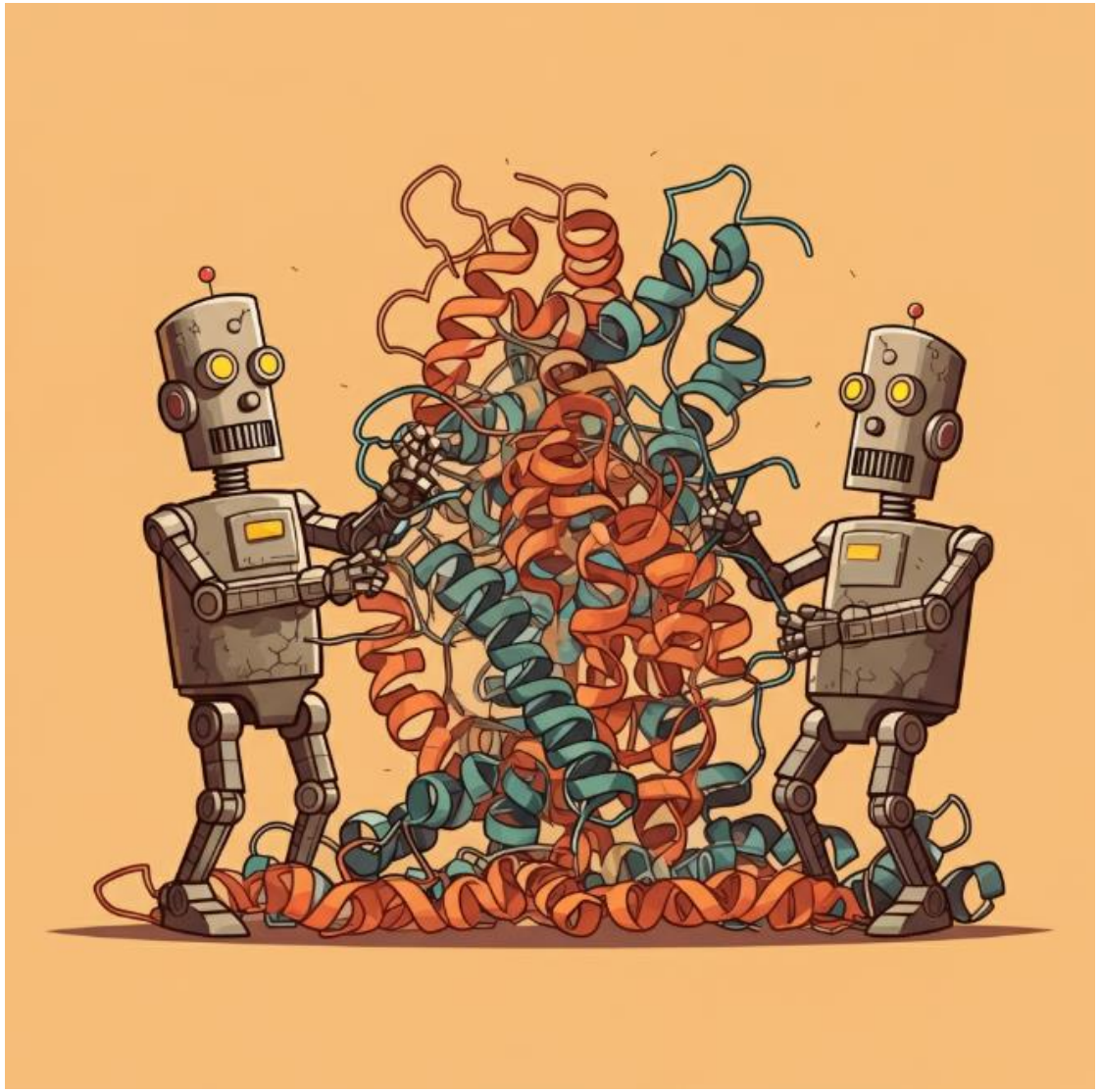




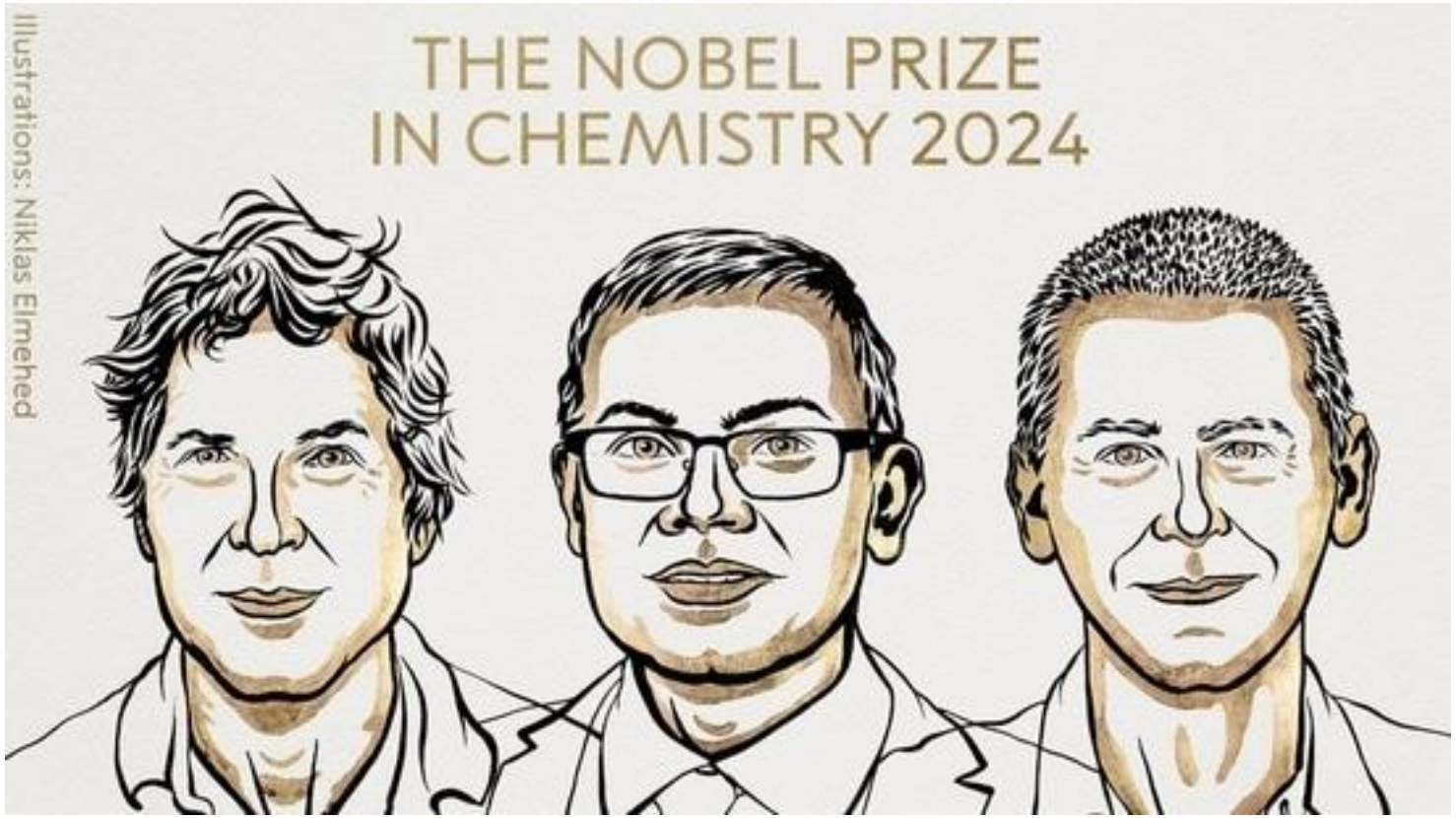




Protein Structure Prediction and Cofolding

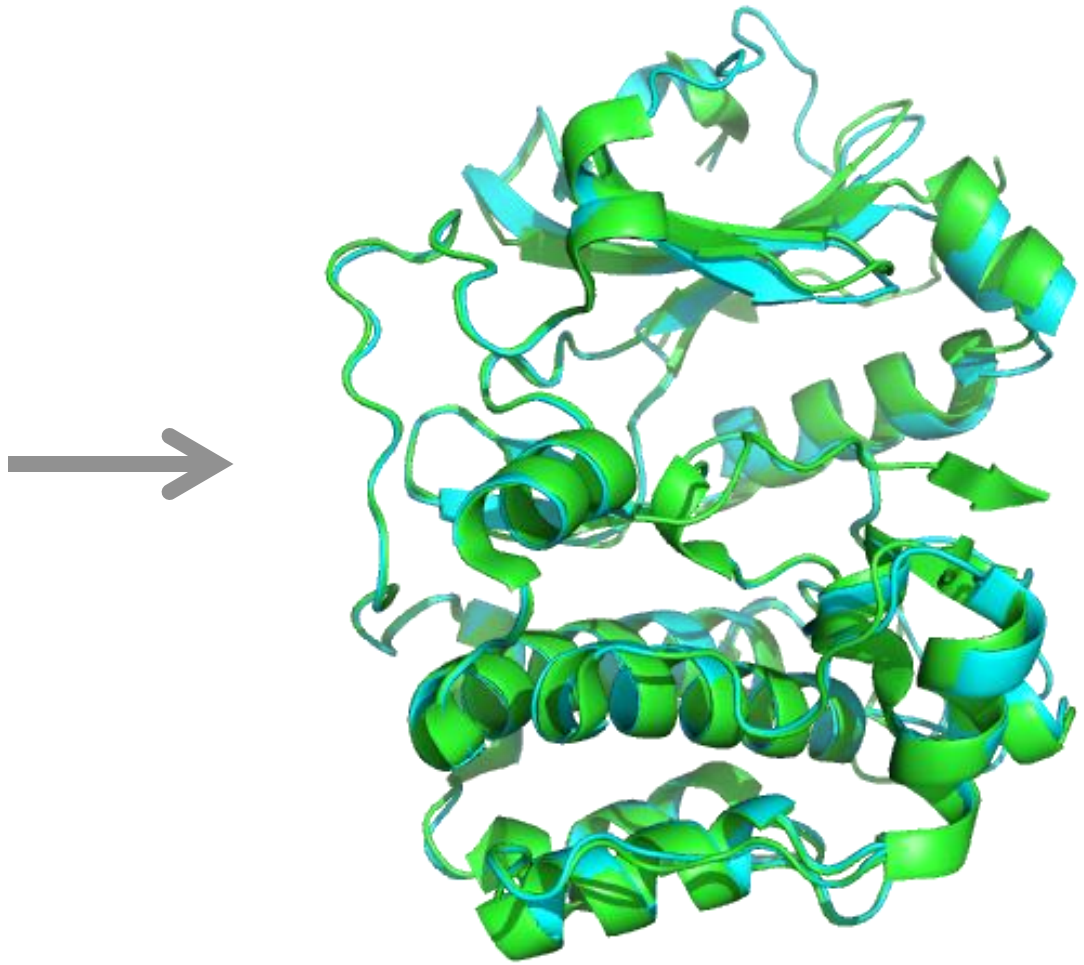


Protein Structure Prediction



Converting Amino Acid Sequence to 3D Structure

VTMNEFEYLKLLGKGTFGKVILVKEKATGRYYAMKI
LKKEVIVAKDEVAHTLTENRVLQNSRHPFLTALKYS
FQTHDRLCFVMEYANGGELFFHLSRERVFSEDRAR
FYGAEIVSALDYLHSEKNVVYRDLKLENLMLDKDG
HIKITDFGLCKEGIKDGATMKFCGTPEYLAPEVLED
NDYGRAVDWWGLGVVMYEMMCGRLPFYNQD
HEKLFELILMEEIRFPRTLGPEAKSLLSGLLKKDPKQ
RLGGGSEDAKEIMQHRFFAGIVWQHVYEKKLSPP
FKPQVTSETDTRYFDEEFTAQRPHFPQFDYSA



AlphaFold and related techniques convert sequence to 3D structure

Can We Use AlphaFold2 Structures For Compound Design or Discovery?

RETURN TO ISSUE | < PREV MACHINE LEARNING AND... NEXT >

Are Deep Learning Structural Models Sufficiently Accurate for Virtual Screening? Application of Docking Algorithms to AlphaFold2 Predicted Structures

Anna M. Díaz-Rovira, Helena Martín, Thijs Beuming, Lucía Díaz, Victor Guallar*, and Soumya S. Ray*

Cite this: *J. Chem. Inf. Model.* 2023, 63, 6, 1668−1674
Publication Date: March 9, 2023
https://doi.org/10.1021/acs.jcim.2c01270
Copyright © 2023 American Chemical Society
Request reuse permissions

Article Views 2994 | Altmetric 38 | Citations 1
LEARN ABOUT THESE METRICS

Share Add to Export RIS

PROTEIN SCIENCE

TOOLS FOR PROTEIN SCIENCE | Open Access

Evaluation of AlphaFold2 structures as docking targets

Matthew Holcomb, Ya-Ting Chang, David S. Goodsell, Stefano Forli

First published: 07 December 2022 | https://doi.org/10.1002/pro.4530 | Citations: 4

Review Editor: Nir Ben-Tal

Funding information National Institute of General Medical Sciences, Grant/Award Number: GM069832

How accurately can one predict drug binding modes using AlphaFold models?

Masha Karelina, Joseph J. Noh, Ron O. Dror

doi: https://doi.org/10.1101/2023.05.18.541346

This is a preprint. It has not been certified by a journal but peer reviews are

0 | 4 | 0 | 0 | 1 | 0 | 105

AlphaFold2 structures template ligand discovery

Jiankun Lyu, Nicholas Kapolka, Ryan Gumpper, Assaf Alon, Liang Wang, Manish K. Jain, Ximena Barros-Álvarez, Kensuke Sakamoto, Yoojoong Kim, Jeffrey DiBerto, Kuglae Kim, Tia A. Tummino, Sijie Huang, John J. Irwin, Olga O. Tarkhanova, Yurii Moroz, Georgios Skiniotis, Andrew C. Kruse, Brian K. Shoichet, Bryan L. Roth

doi: https://doi.org/10.1101/2023.12.20.572662

This article is a preprint and has not been certified by peer review [what does this mean?]

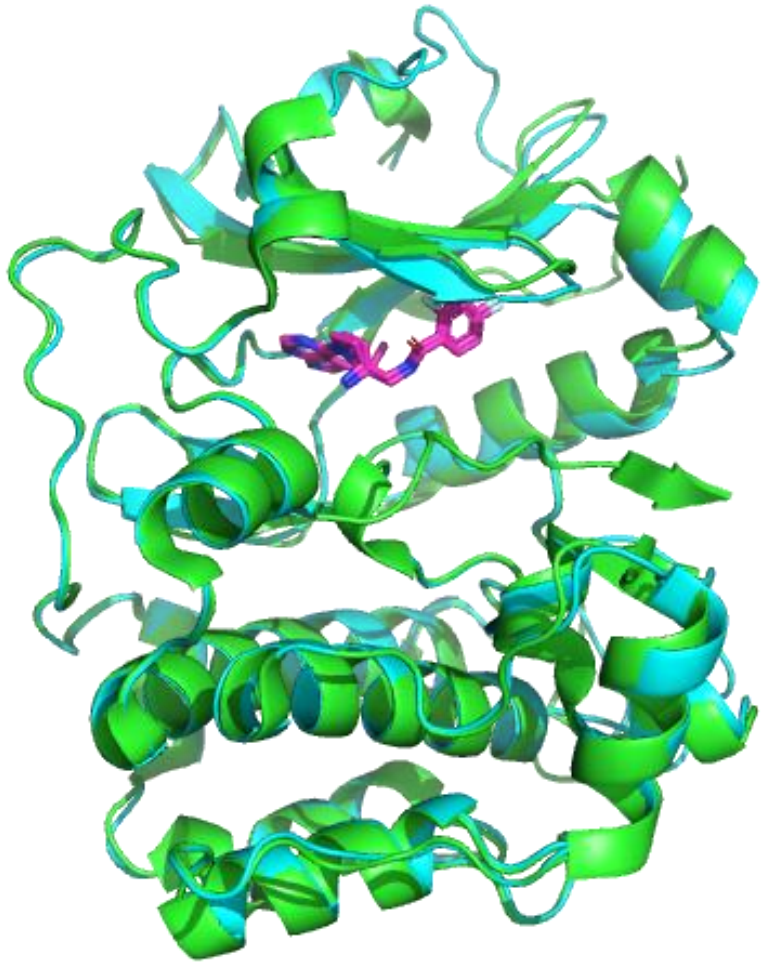
0 | 0 | 0 | 1 | 1 | 0 | 139

Performance of AF2 structures for docking is roughly equivalent to that of apo structures

Co-Folding Enables the Prediction of Ligand Poses With Protein Structures

VTMNEFEYLKLLGKGTFGKVILVKEKATGRYYAMKI
LKKEVIVAKDEVAHTLTENRVLQNSRHPFLTALKYS
FQTHDRLCFVMEYANGGELFFHLSRERVFSEDRAR
FYGAEIVSALDYLHSEKNVVYRDLKLENLMLDKDG
HIKITDFGLCKEGIKDGATMKFCGTPEYLAPEVLED
NDYGRAVDWWGLGVVMYEMMCGRLPFYNQD
HEKLFELILMEEIRFPRTLGPEAKSLLSGLLKKDPKQ
RLGGGSEDAKEIMQHRFFAGIVWQHVYEKKLSPP
FKPQVTSETDTRYFDEEFTAQRPHFPQFDYSA

CCc1c[nH]c2c1c(ncn2)N3CCC(C3)(CNC(=O)c4ccc(cc4F)F)N



Co-folding was introduced with DragonFold, AlphaFold 3, and RoseTTaFold All Atom

Attack of the (AlphaFold3) Clones

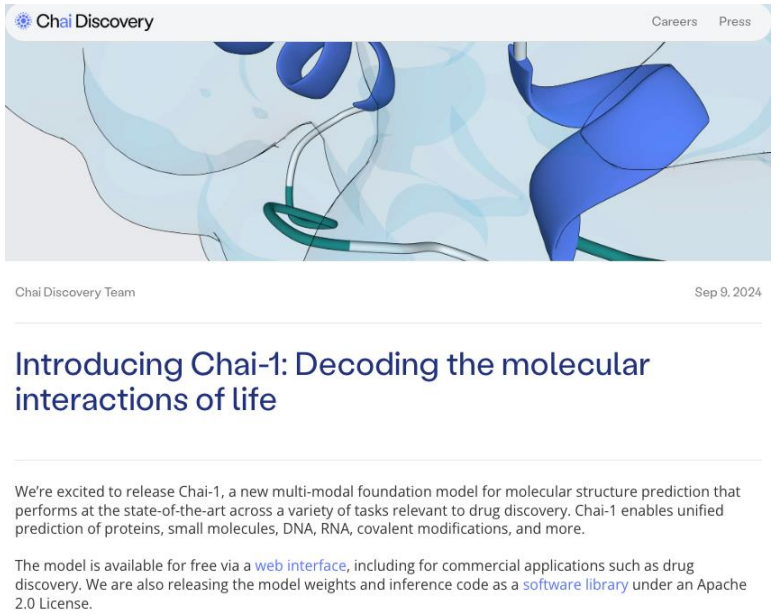
RELAY THERAPEUTICS

OpenFold

open fold
Democratizing AI for Biology

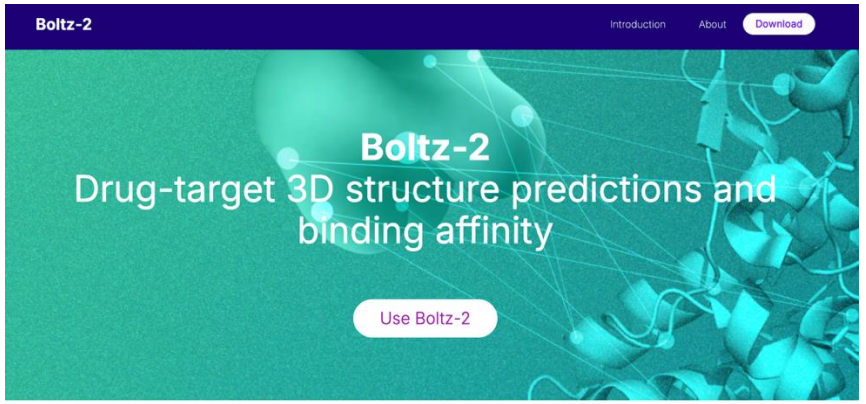
https://openfold.io/

Chai-1



https://www.chaidiscovery.com/blog/introducing-chai-1

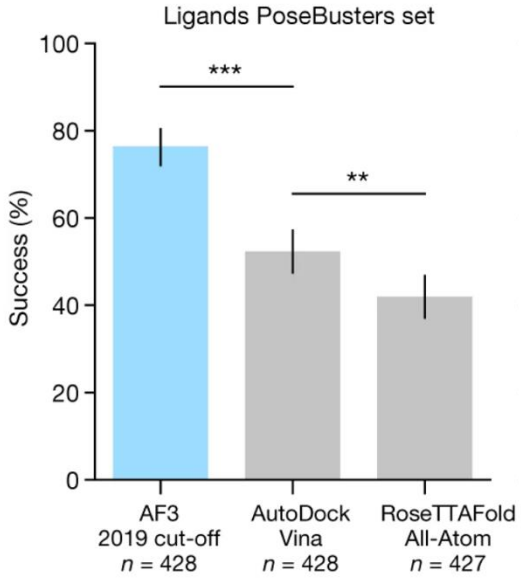
Boltz-2



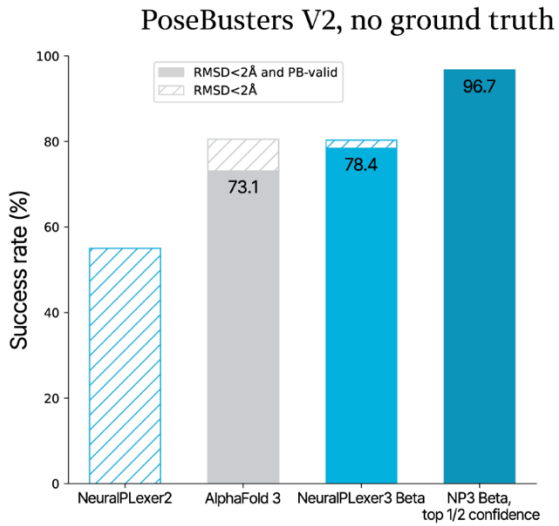
https://www.rxrx.ai/boltz-2

AlphaFold3 license restrictions motivated the development of programs with similar functionality

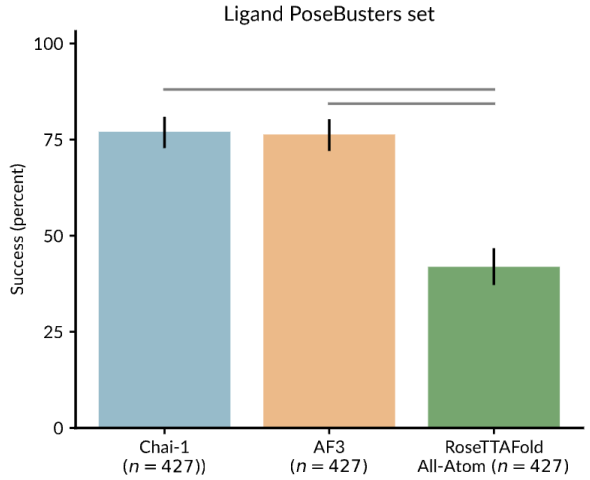
PoseBusters Has Become a Standard Benchmark for Co-folding Methods



AlphaFold 3



NeuralPlexer

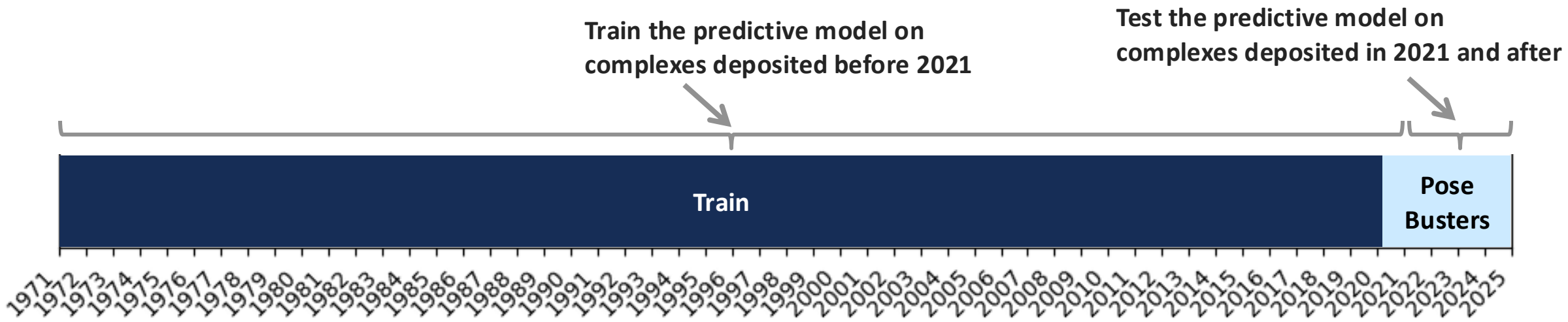


Chai-1

Success rates for co-folding benchmarks may not reflect real-world performance

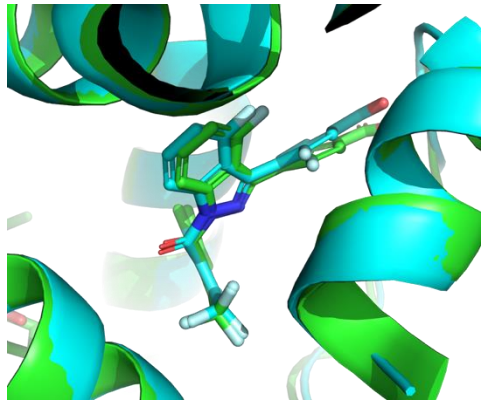
Buttenschoen, M., Morris, G. M., & Deane, C. M. (2024). PoseBusters: AI-based docking methods fail to generate physically valid poses or generalise to novel sequences. *Chemical Science*, *15*(9), 3130-3139.

Time Splits Allow Training Data to Leak Into Predictions

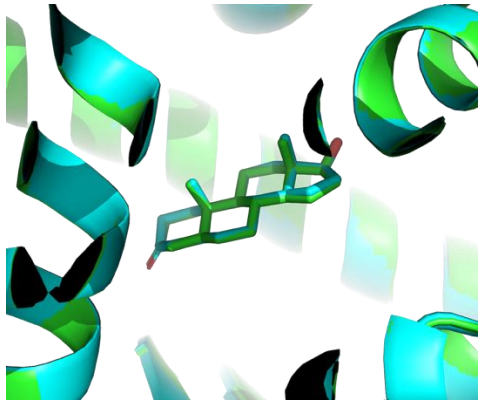


Many test set complexes may have close analogs in the training set

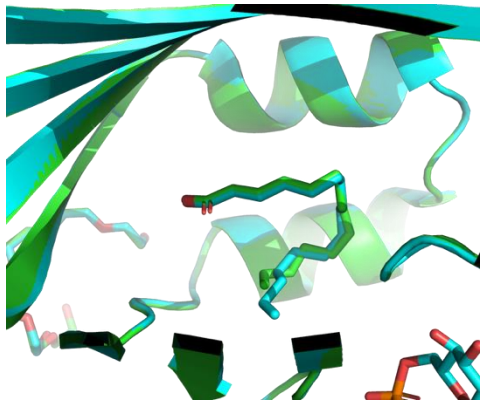
Time Splits Can Lead to Significant Train / Test Overlap



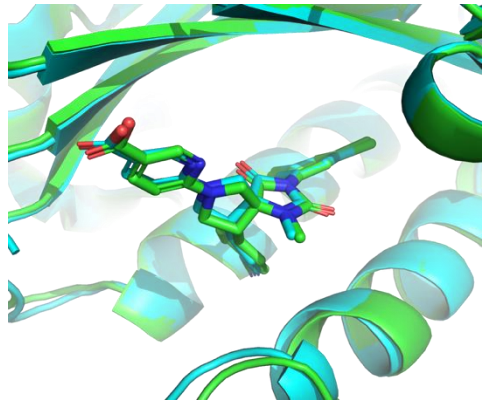
8fav 4Y5 2023 | 5c4s 4Y5 2015



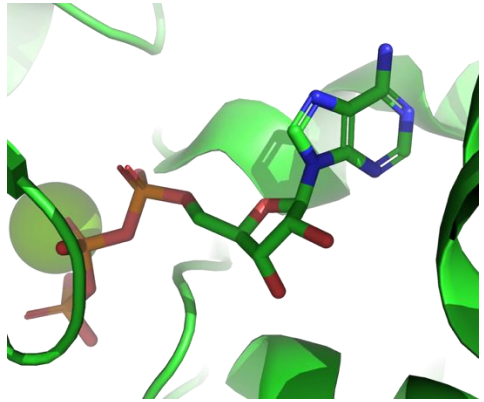
7zu2 DHT 2023 | 2ama DHT 2006



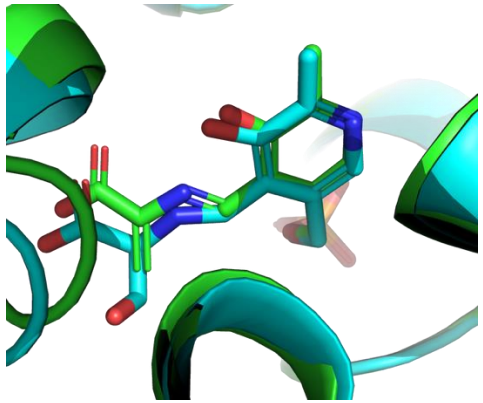
7wpw F15 2023 | 4tkj PLM 2015



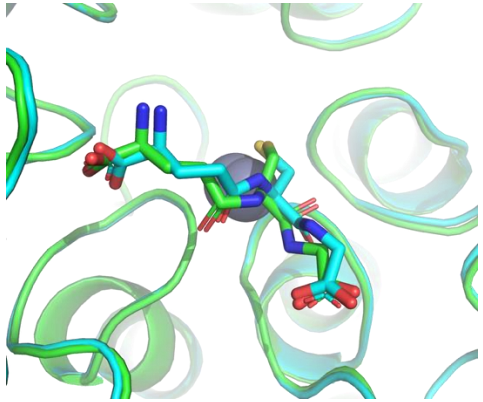
7kc5 BJZ 2023 | 3m6f BJZ 2010



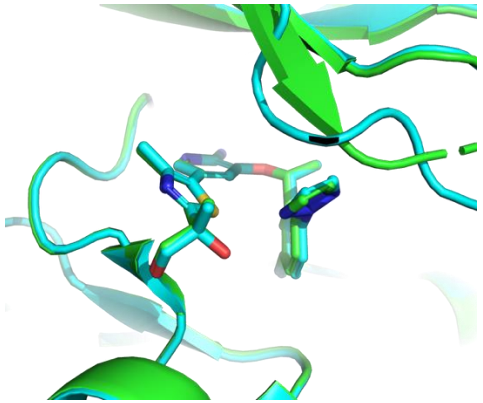
7kru ATP 2021 | 4jne ATP 2013



7l03 F9F 2022 | 2cll F9F 2007



8d19 GSH 2023 | 3e73 GSH 2009



7r7r AWJ 2022 | 4cdo AWJ 2014

The Most Popular Slide at the 2025 CADD GRC

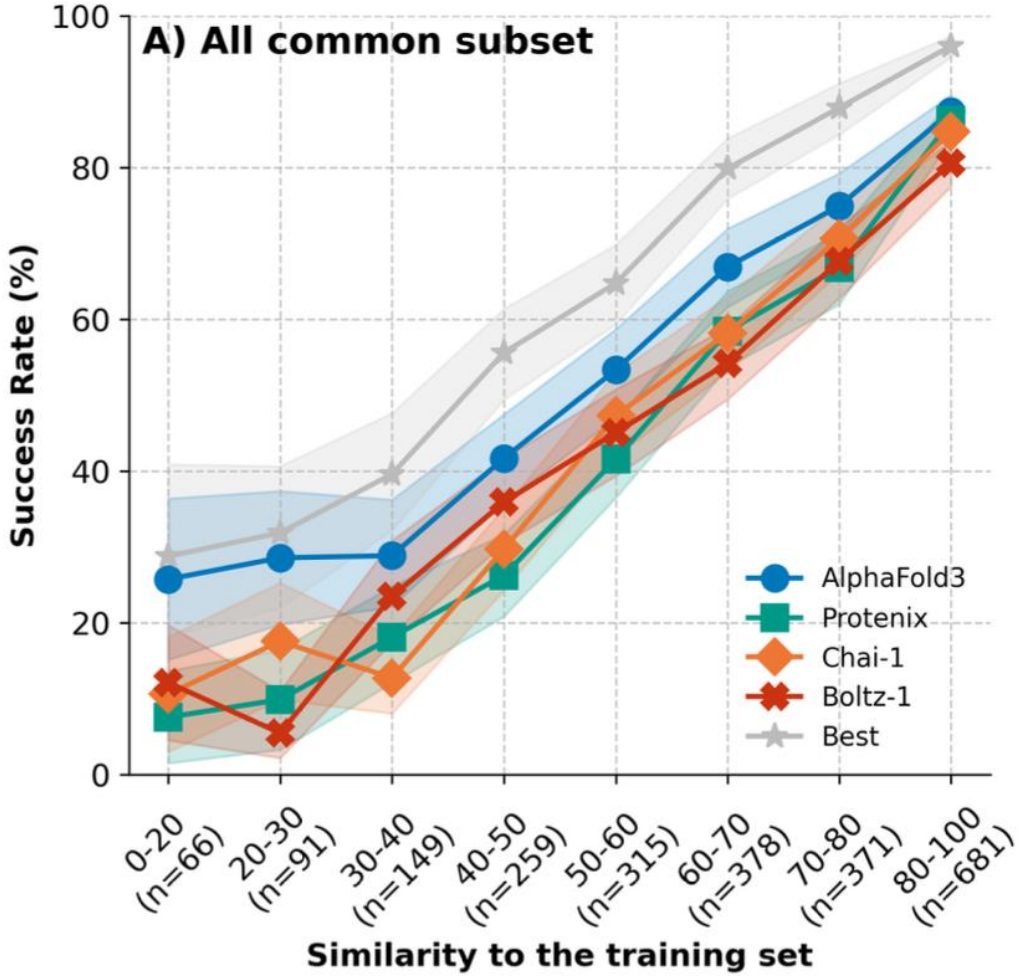
HAVE PROTEIN-LIGAND CO-FOLDING METHODS MOVED BEYOND MEMORISATION?

Peter Škrinjar
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
peter.skrinjar@unibas.ch

Jérôme Eberhardt
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
jerome.eberhardt@unibas.ch

Janani Durairaj
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
janani.durairaj@unibas.ch

Torsten Schwede
Biozentrum, University of Basel
SIB Swiss Institute of Bioinformatics
torsten.schwede@unibas.ch



Co-folding success rate correlates with training set similarity

Benchmarking Co-folding at Orthosteric and Allosteric Sites

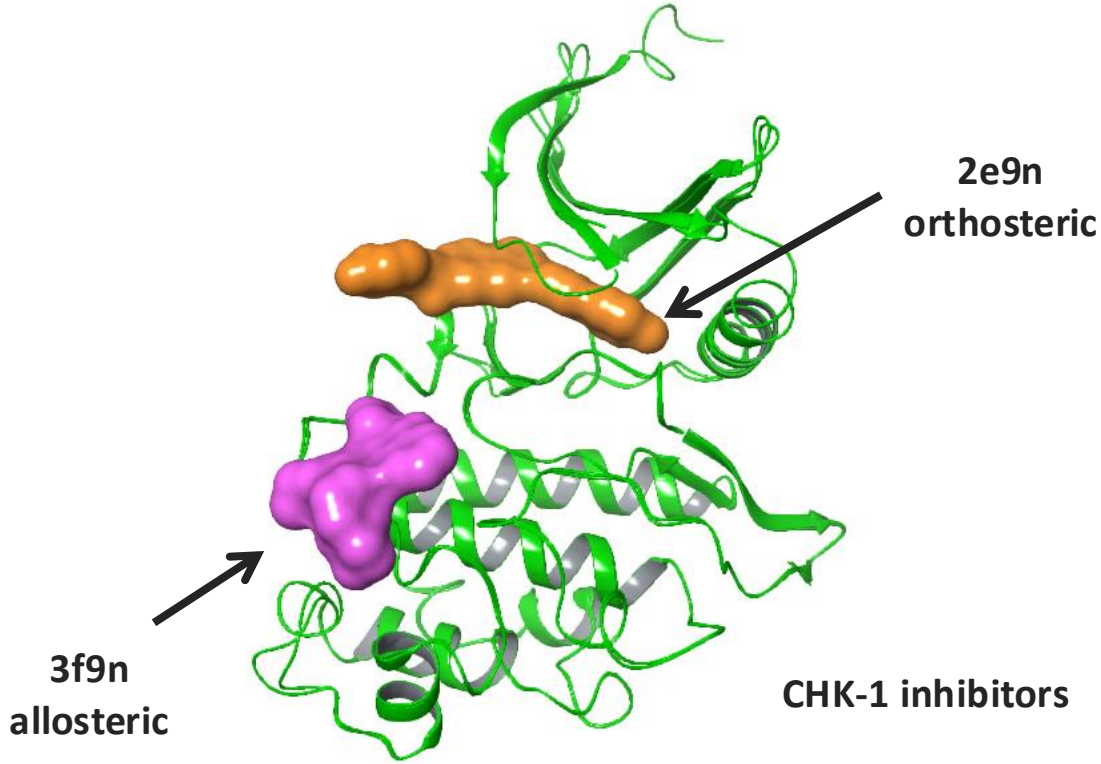
Article

Co-folding, the Future of Docking – Prediction of Allosteric and Orthosteric Ligands

Eva Nittinger, Özge Yoluk, Alessandro Tibo, Gustav Olanders, Christian Tyrchan

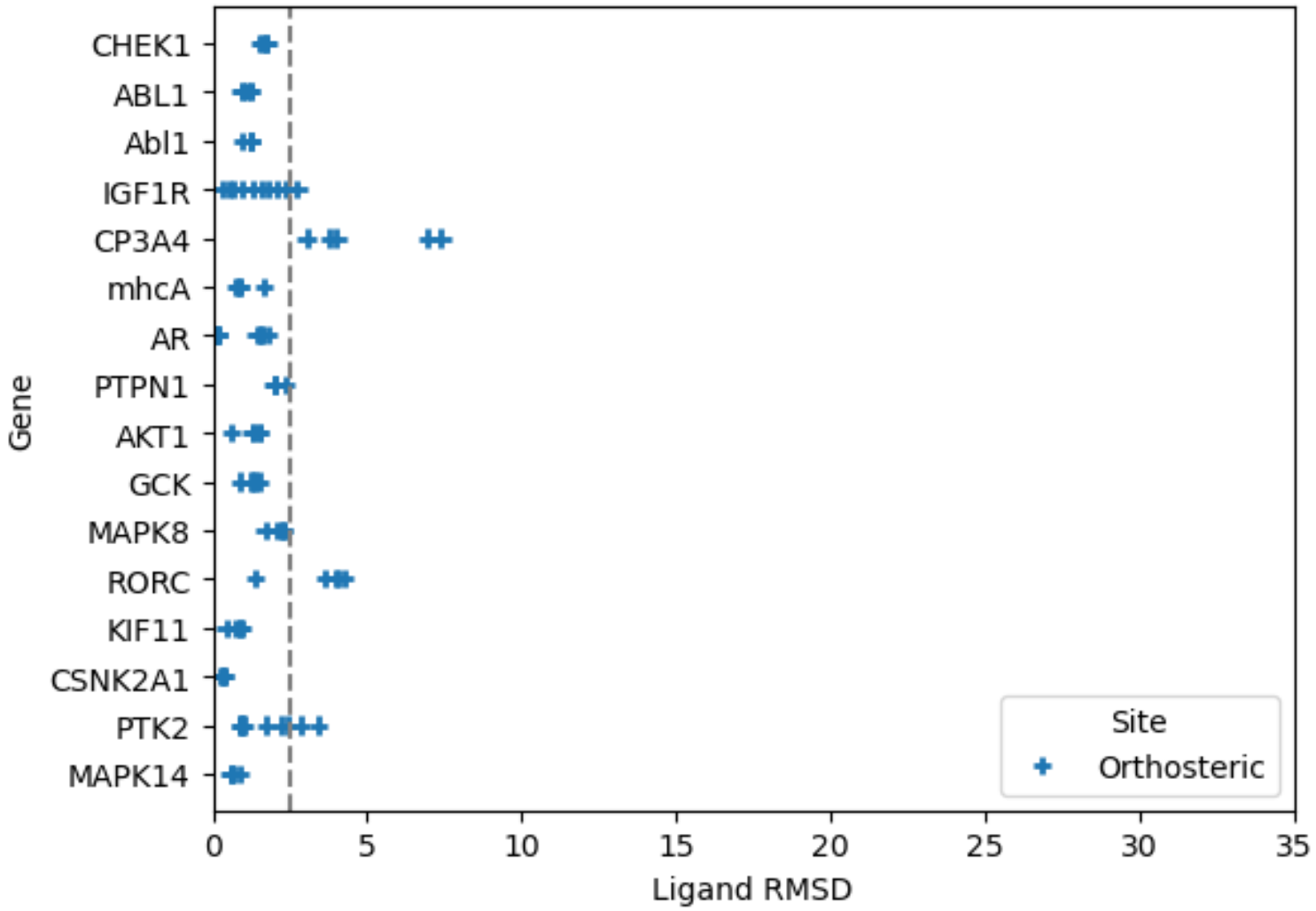
This is a preprint; it has not been peer reviewed by a journal.

https://doi.org/10.21203/rs.3.rs-6526650/v1



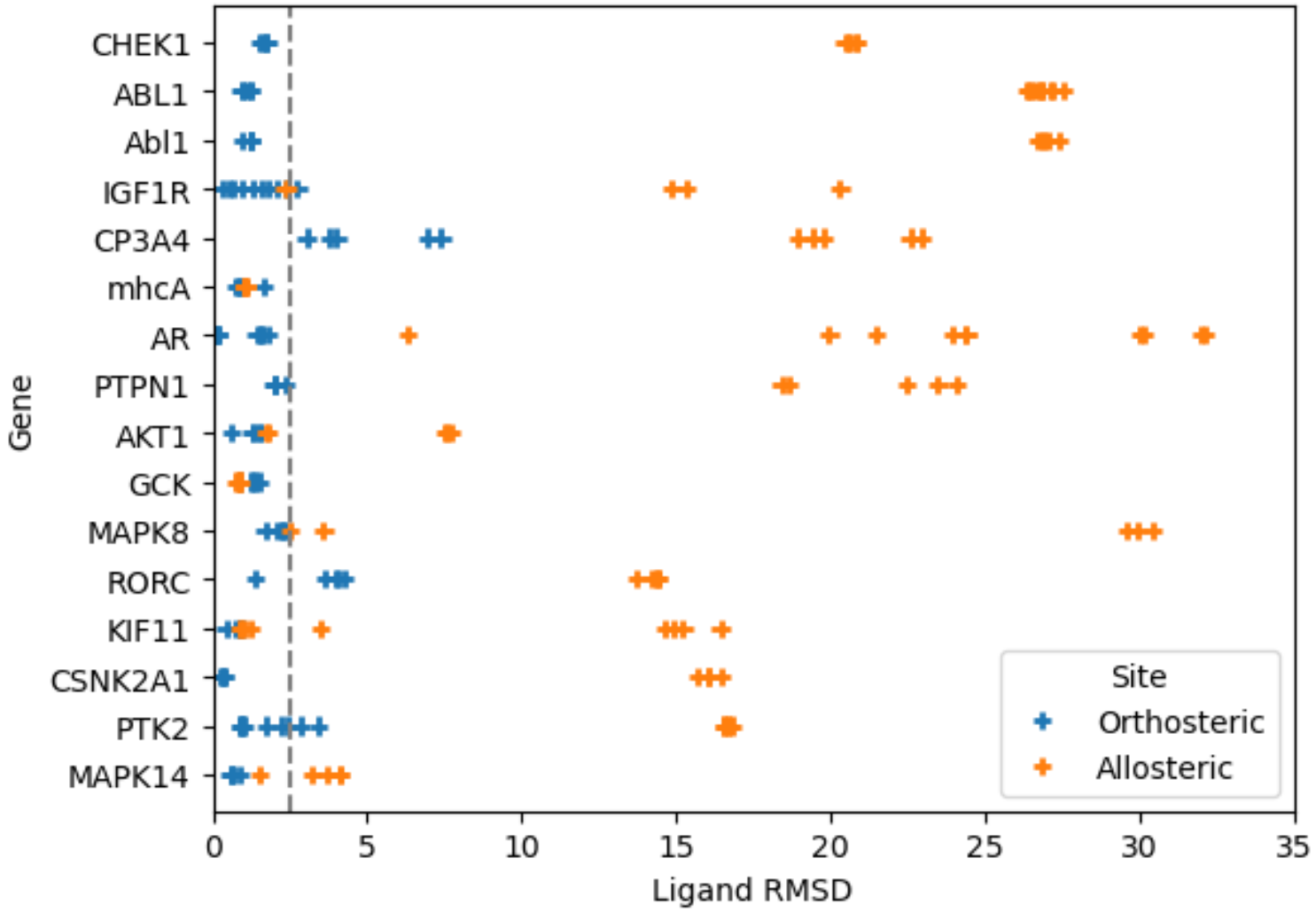
16 pairs protein/ligand complex with both allosteric and orthosteric binders

Co-Folding Performs Well With Orthosteric Binding Sites



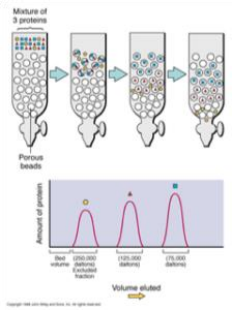
Study based on https://www.researchsquare.com/article/rs-6526650/v1

Co-Folding Performs Poorly With Allosteric Binding Sites



Study based on https://www.researchsquare.com/article/rs-6526650/v1

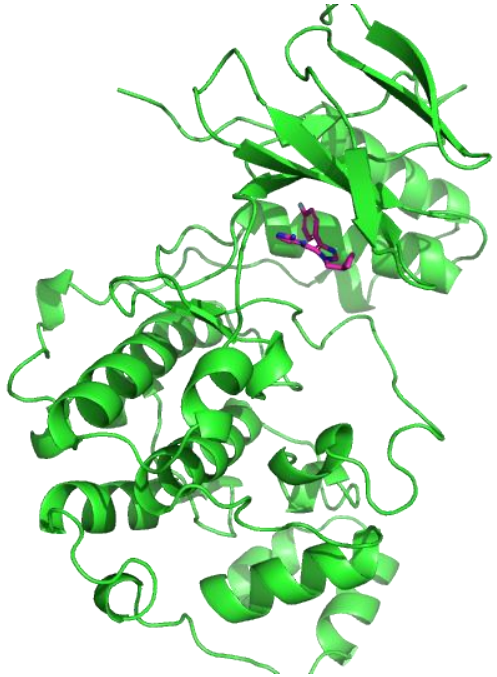
Co-Folding Generates Hypotheses to Bridge Experiments



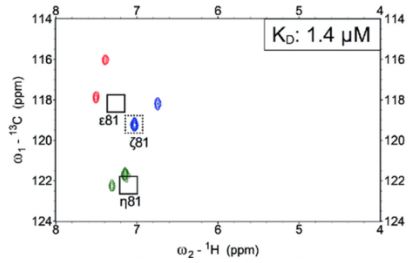
SEC-MS

HTS

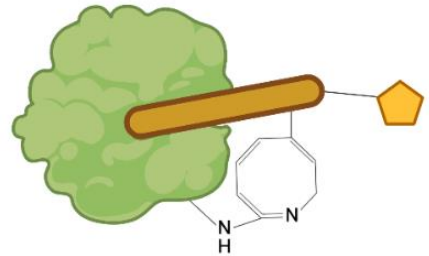
DEL



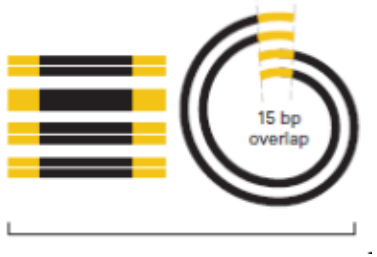
HDX MS



NMR

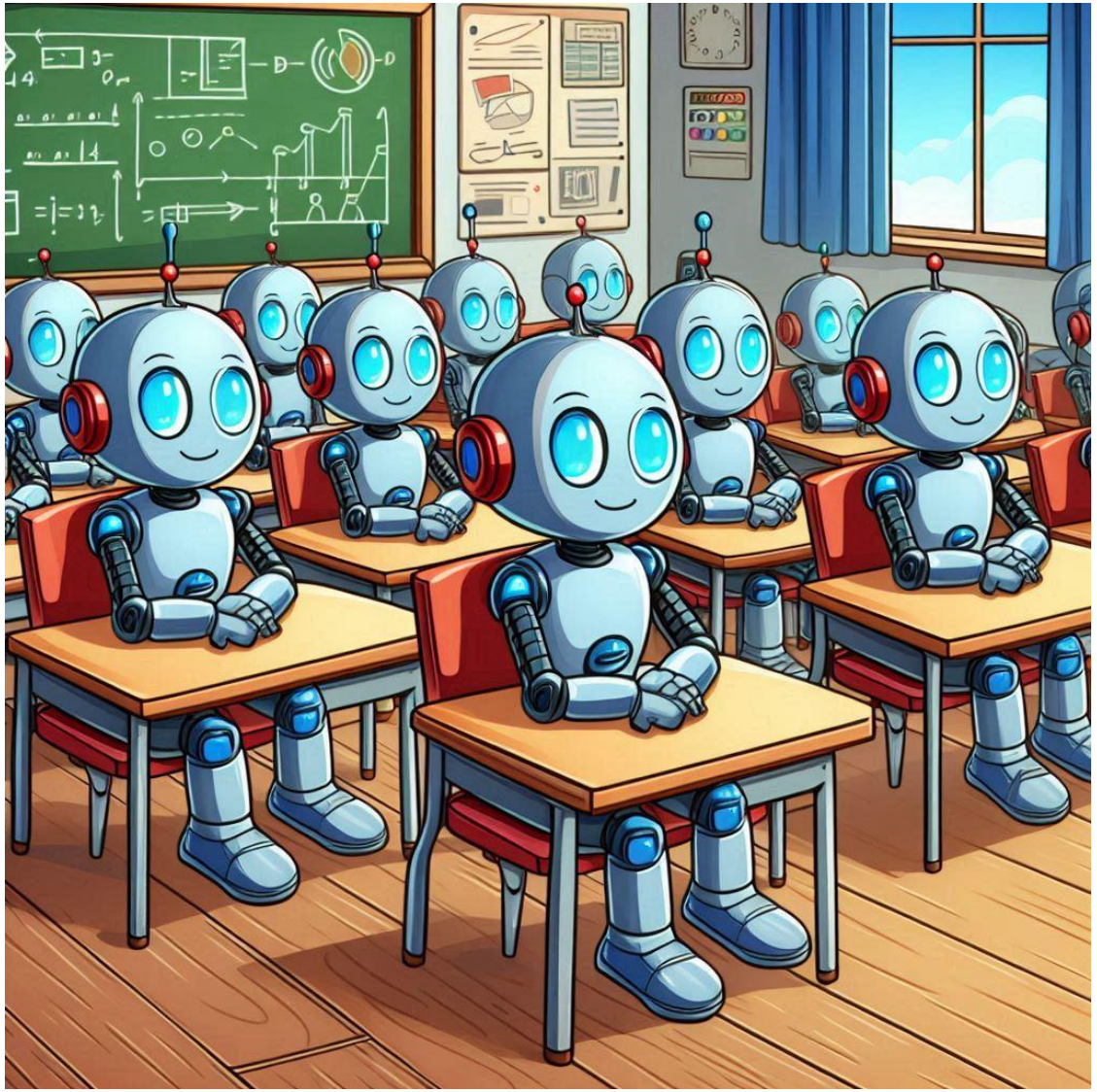


Photoaffinity Probes

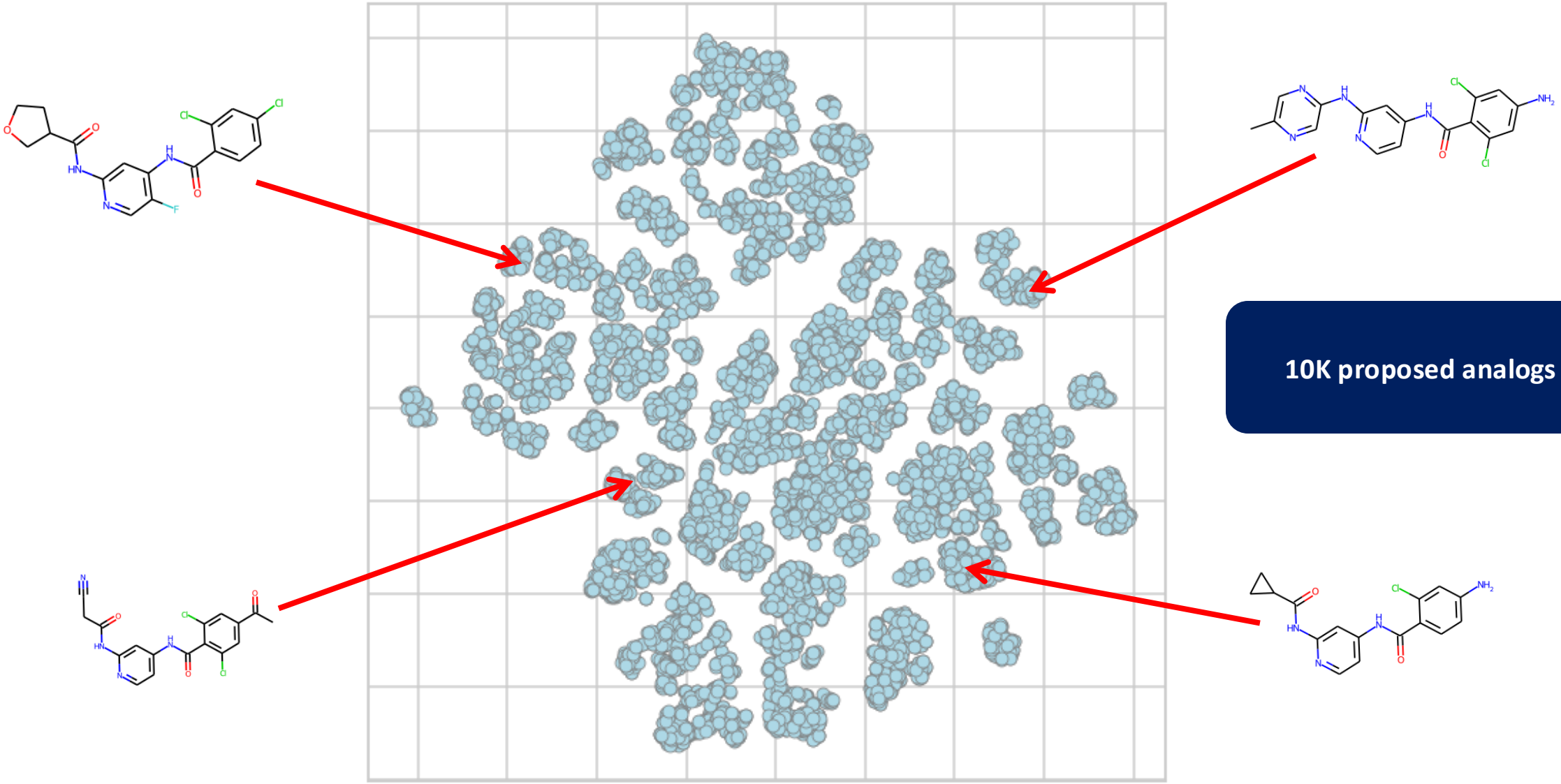


Site-Directed Mutegenesis

Active Learning



Prioritizing Molecules For Synthesis

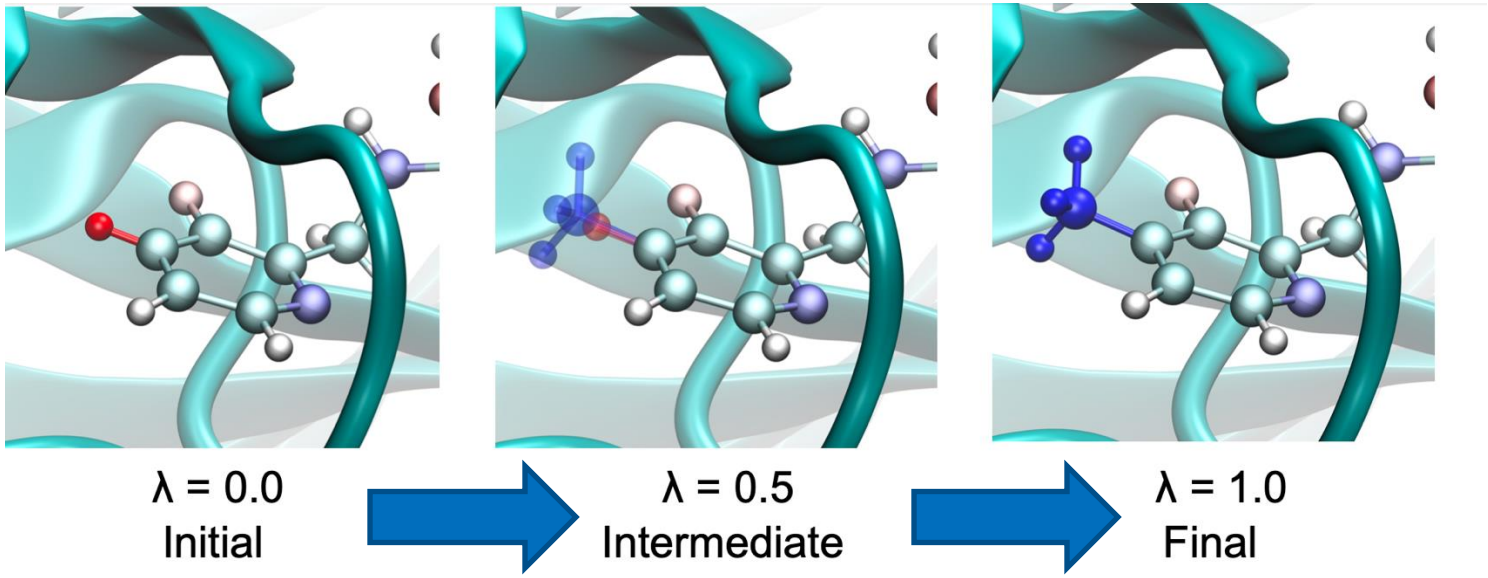


10K proposed analogs

Free Energy Perturbation (FEP): Transformation

FEP evaluates binding energy by morphing a known molecule into a new molecule

1. **Alchemically morph one molecule to another**
2. **Run simulations using Molecular Dynamics**
3. **Compute free energy of transforming the molecule**



Calculations typically take 4-8 hrs/molecule
10K calculations would require ~9 GPU years

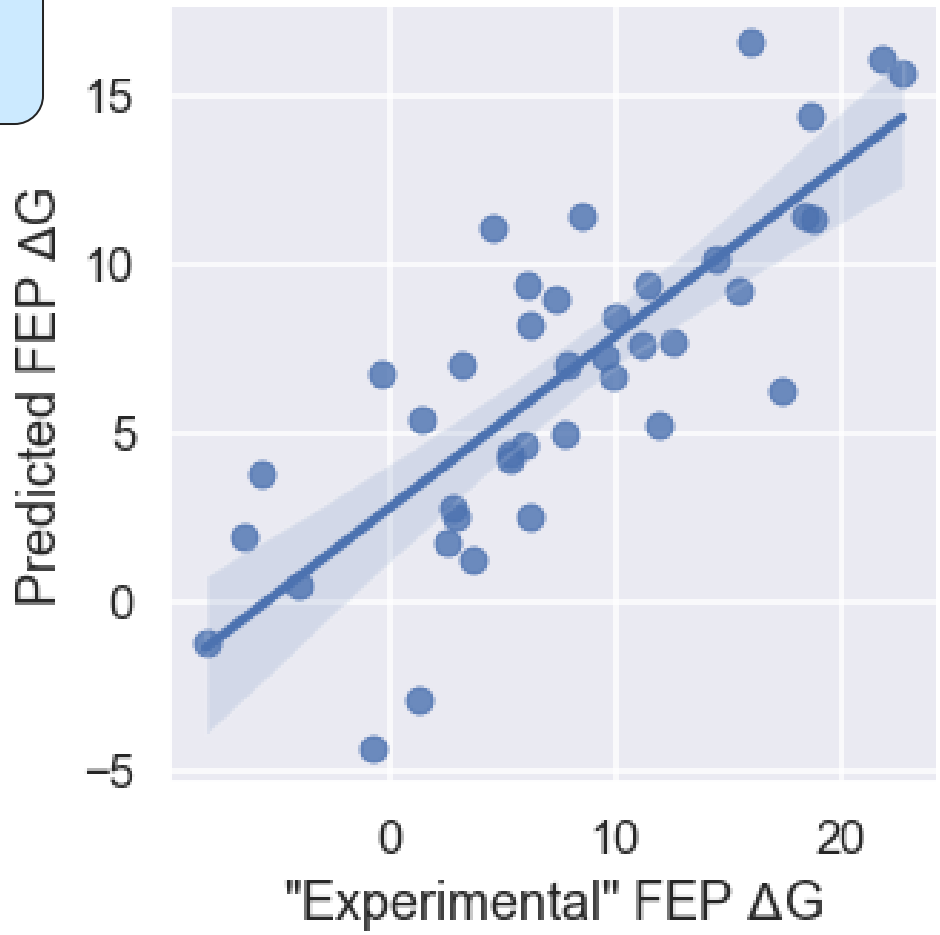
Build a Machine Learning Model to Predict FEP ΔG

Time to process 10K molecules

Method	Time (sec)
FEP	2.9×10^8
Machine Learning Model	7.5

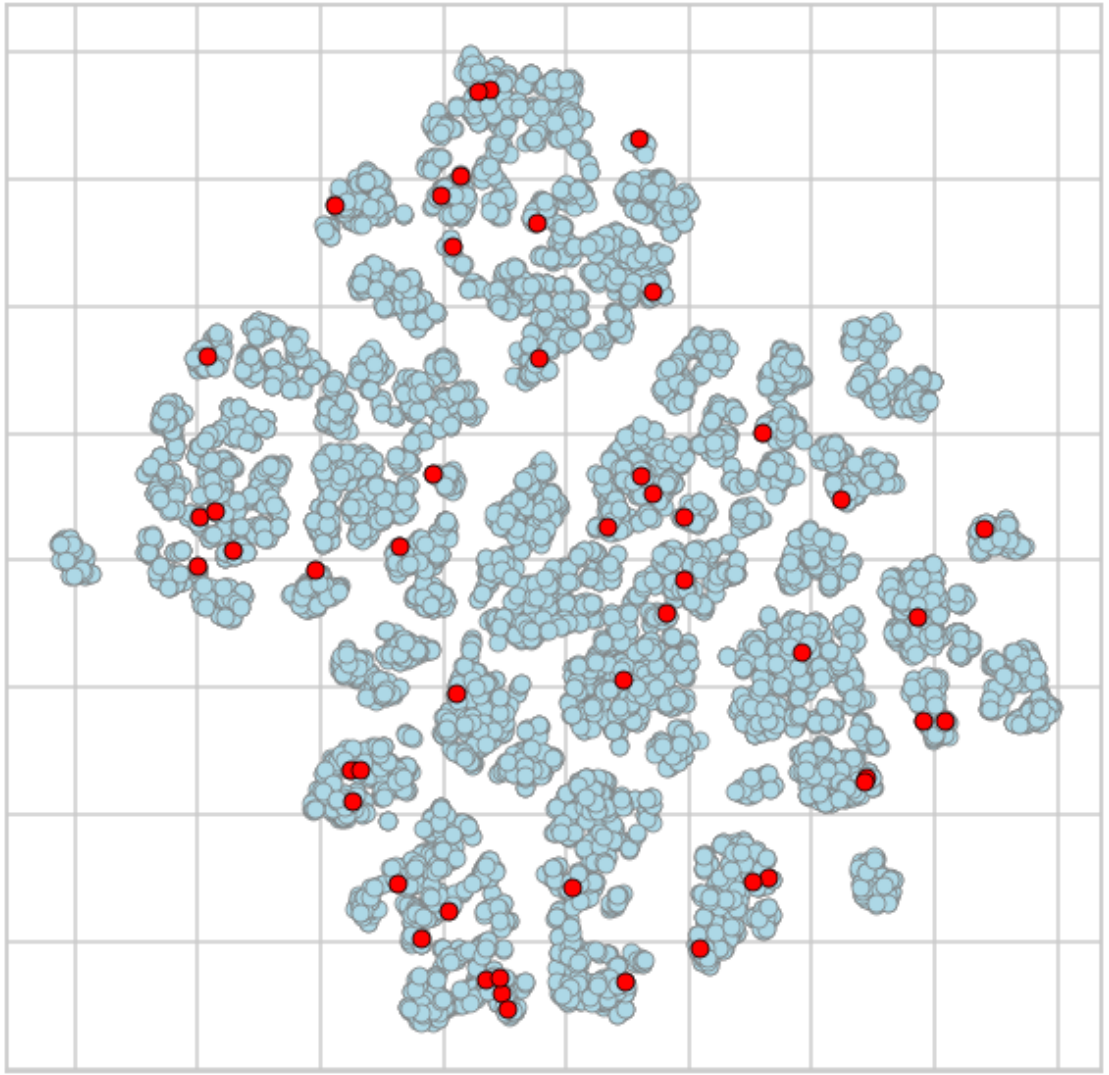
9 GPU years

Machine learning model is 39 million times faster than FEP*



***Training on 100 molecules + inference on 10K molecules**

Sample 50 of 10K molecules and Run FEP



FEP

	SMILES	dG bind
112	COc1c(F)cccc1C(=O)Nc1cc(NC(=O)C2COC2)ncc1F	-12.83
275	CN(C)c1nccc(Nc2cc(NC(=O)c3c(F)ccc(F)c3Cl)ccn2)n1	-10.16
563	CCNC(=O)Nc1cc(NC(=O)c2cccc(O)c2Cl)ccn1	-7.68
589	O=C(CC1CC1)Nc1cc(NC(=O)c2cc(O)ccc2Cl)c(F)cn1	-7.51
1170	COC(=O)c1cccc(Nc2cc(NC(=O)c3ccccc3Cl)c(F)cn2)n1	-4.58
1342	CNC(=O)Nc1cc(NC(=O)c2c(F)cccc2OC)ccn1	-3.84
1657	CC(C)NC(=O)Nc1cc(NC(=O)c2c(F)ccc(F)c2Cl)ccn1	-2.71
1664	Cc1cccc(Cl)c1C(=O)Nc1cc(Nc2ccncn2)ncc1F	-2.69
1974	O=C(Nc1cc(NC(=O)C2CCCO2)ncc1F)c1c(F)cccc1Br	-1.80
2071	O=C(Nc1cc(NC(=O)c2c(Cl)cccc2Br)ccn1)NC1CC1	-1.51
2372	Cc1cc(Cl)c(C(=O)Nc2ccnc(Nc3cc(C)nc(CO)n3)c2)c(...	-0.74
2542	Cc1cc(Nc2cc(NC(=O)c3ccccc3C)ccn2)nc(NC2CC2)n1	-0.32
2792	CSCC(=O)Nc1cc(NC(=O)c2c(Cl)cc(N)cc2Cl)c(F)cn1	0.31
2815	COC(=O)Nc1cc(NC(=O)c2cccc(F)c2Cl)c(F)cn1	0.37
2880	COc1ccc(C(=O)Nc2cc(NC(=O)C3CC3(F)F)ncc2F)c(Cl)c1	0.54
3265	COc1ccc(Cl)c(C(=O)Nc2ccnc(NC(=O)C3CC3CO)c2)c1	1.45
3355	O=C(Nc1ccnc(NC(=O)C2CC2(F)F)c1)c1ccc(O)cc1F	1.65
3568	CC1CC1C(=O)Nc1cc(NC(=O)c2c(F)cc(F)cc2F)ccn1	2.19
3829	O=C(Nc1cc(Nc2ccncn2)ncc1F)c1cccc(Cl)c1	2.76
3857	COc1ccc(F)c(C(=O)Nc2cc(NC(=O)C3CC3C)ncc2F)c1	2.82
4644	COc1cccc(C(=O)Nc2cc(NC(=O)C3CC(=O)C3)ncc2F)c1F	4.45
4729	Cc1nc(Cl)cc(Nc2cc(NC(=O)c3ccc(N)cc3)ccn2)n1	4.65
4901	CN(C)C(=O)c1cccc(Nc2cc(NC(=O)c3cccc(N)c3Cl)ccn...	5.00
4987	COc1ccc(Nc2cc(NC(=O)c3c(Cl)cccc3Br)ccn2)nc1	5.13

etc.

Generate Molecular Descriptors From Chemical Structures

SMILES

	SMILES
112	COc1c(F)cccc1C(=O)Nc1cc(NC(=O)C2COC2)ncc1F
275	CN(C)c1nccc(Nc2cc(NC(=O)c3c(F)ccc(F)c3Cl)ccn2)n1
563	CCNC(=O)Nc1cc(NC(=O)c2cccc(O)c2Cl)ccn1
589	O=C(CC1CC1)Nc1cc(NC(=O)c2cc(O)ccc2Cl)c(F)cn1
1170	COC(=O)c1cccc(Nc2cc(NC(=O)c3ccccc3Cl)c(F)cn2)n1
1342	CNC(=O)Nc1cc(NC(=O)c2c(F)cccc2OC)ccn1
1657	CC(C)NC(=O)Nc1cc(NC(=O)c2c(F)ccc(F)c2Cl)ccn1
1664	Cc1cccc(Cl)c1C(=O)Nc1cc(Nc2ccncn2)ncc1F
1974	O=C(Nc1cc(NC(=O)C2CCCO2)ncc1F)c1c(F)cccc1Br
2071	O=C(Nc1cc(NC(=O)c2c(Cl)cccc2Br)ccn1)NC1CC1
2372	Cc1cc(Cl)c(C(=O)Nc2ccnc(Nc3cc(C)nc(CO)n3)c2)c(...
2542	Cc1cc(Nc2cc(NC(=O)c3ccccc3C)ccn2)nc(NC2CC2)n1
2792	CSCC(=O)Nc1cc(NC(=O)c2c(Cl)cc(N)cc2Cl)c(F)cn1
2815	COC(=O)Nc1cc(NC(=O)c2cccc(F)c2Cl)c(F)cn1
2880	COc1ccc(C(=O)Nc2cc(NC(=O)C3CC3(F)F)ncc2F)c(Cl)c1
3265	COc1ccc(Cl)c(C(=O)Nc2ccnc(NC(=O)C3CC3CO)c2)c1
3355	O=C(Nc1ccnc(NC(=O)C2CC2(F)F)c1)c1ccc(O)cc1F
3568	CC1CC1C(=O)Nc1cc(NC(=O)c2c(F)cc(F)cc2F)ccn1
3829	O=C(Nc1cc(Nc2ccncn2)ncc1F)c1cccc(Cl)c1
3857	COc1ccc(F)c(C(=O)Nc2cc(NC(=O)C3CC3C)ncc2F)c1

Molecular Descriptors

0	1	2	3	4	5	6	7	8	9	...	90	91	92	93	94	95	96	97	98	99
2.266	-2.240	2.246	-2.381	6.062	-0.133	19.144	10.089	1.784	855.897	...	23.320	0.0	4.984	5.918	0.000	30.957	47.660	0.0	5.75	15.370
2.192	-2.121	2.244	-2.152	6.342	0.102	35.496	10.155	1.797	1030.708	...	40.780	0.0	14.952	0.000	0.000	29.629	64.946	0.0	0.00	15.533
2.133	-2.093	2.269	-2.281	6.354	0.102	35.496	10.165	2.122	736.412	...	35.044	0.0	10.301	0.000	6.924	17.178	47.115	0.0	5.75	15.950
2.243	-2.100	2.322	-2.159	6.341	-0.116	35.496	10.163	1.742	839.883	...	34.921	0.0	4.984	5.918	19.262	10.634	46.866	0.0	5.75	10.634
2.139	-2.103	2.221	-2.138	6.340	0.059	35.496	10.177	1.811	1045.192	...	40.801	0.0	9.968	0.000	0.000	17.743	76.825	0.0	0.00	10.634
2.175	-2.107	2.225	-2.246	6.062	0.102	19.142	10.142	2.210	736.412	...	23.444	0.0	10.301	0.000	0.000	24.791	47.909	0.0	5.75	20.687
2.193	-2.112	2.239	-2.339	6.342	0.102	35.496	10.155	2.159	814.717	...	35.044	0.0	10.301	0.000	19.889	10.634	52.683	0.0	0.00	15.950
2.155	-2.112	2.235	-2.151	6.342	0.102	35.496	10.087	1.847	899.917	...	34.831	0.0	14.952	0.000	6.924	10.634	71.020	0.0	0.00	10.634
2.241	-2.121	2.233	-2.270	9.103	-0.124	79.919	10.164	1.771	836.892	...	39.250	0.0	4.984	0.000	18.946	17.240	52.133	0.0	0.00	10.634
2.229	-2.101	2.257	-2.303	9.103	0.102	79.919	10.177	1.693	775.147	...	50.974	0.0	10.301	0.000	18.883	10.634	51.587	0.0	0.00	15.950
2.158	-2.111	2.274	-2.134	6.401	0.102	35.497	10.124	1.804	1019.578	...	46.432	0.0	14.952	0.000	20.454	10.634	69.219	0.0	0.00	10.634
2.218	-2.108	2.243	-2.132	6.049	0.102	16.149	10.091	1.486	1015.283	...	29.179	0.0	14.952	0.000	32.731	15.950	65.482	0.0	0.00	15.950
2.178	-2.111	2.288	-2.171	7.988	-0.113	35.497	10.154	2.248	812.169	...	63.971	0.0	4.984	0.000	0.000	28.376	45.822	0.0	0.00	16.367
2.159	-2.095	2.225	-2.159	6.341	0.102	35.496	10.168	2.248	770.928	...	35.107	0.0	4.984	0.000	0.000	17.743	52.683	0.0	0.00	10.634
2.609	-2.104	2.486	-2.189	6.342	-0.119	35.496	10.165	1.745	923.537	...	34.921	0.0	4.984	5.918	12.343	17.743	46.866	0.0	5.75	15.370
2.419	-2.098	2.382	-2.170	6.341	-0.117	35.496	10.163	1.685	842.936	...	34.921	0.0	4.984	11.836	6.421	24.350	47.115	0.0	5.75	15.370
2.609	-2.085	2.485	-2.185	6.045	-0.119	19.287	10.157	1.690	857.514	...	23.320	0.0	4.984	5.918	12.343	10.634	47.909	0.0	5.75	10.634
2.387	-2.090	2.379	-2.176	6.046	-0.117	19.149	10.146	1.737	833.777	...	23.320	0.0	4.984	11.836	13.345	10.634	53.477	0.0	0.00	10.634
2.113	-2.093	2.204	-2.132	6.306	0.102	35.496	10.179	1.800	875.352	...	34.831	0.0	14.952	0.000	0.000	10.634	71.523	0.0	0.00	10.634
2.388	-2.102	2.380	-2.175	6.048	-0.117	19.144	10.152	1.773	872.755	...	23.320	0.0	4.984	11.836	13.345	17.743	47.660	0.0	5.75	15.370

Build a Machine Learning Models to Predict FEP ΔG From Molecular Descriptors

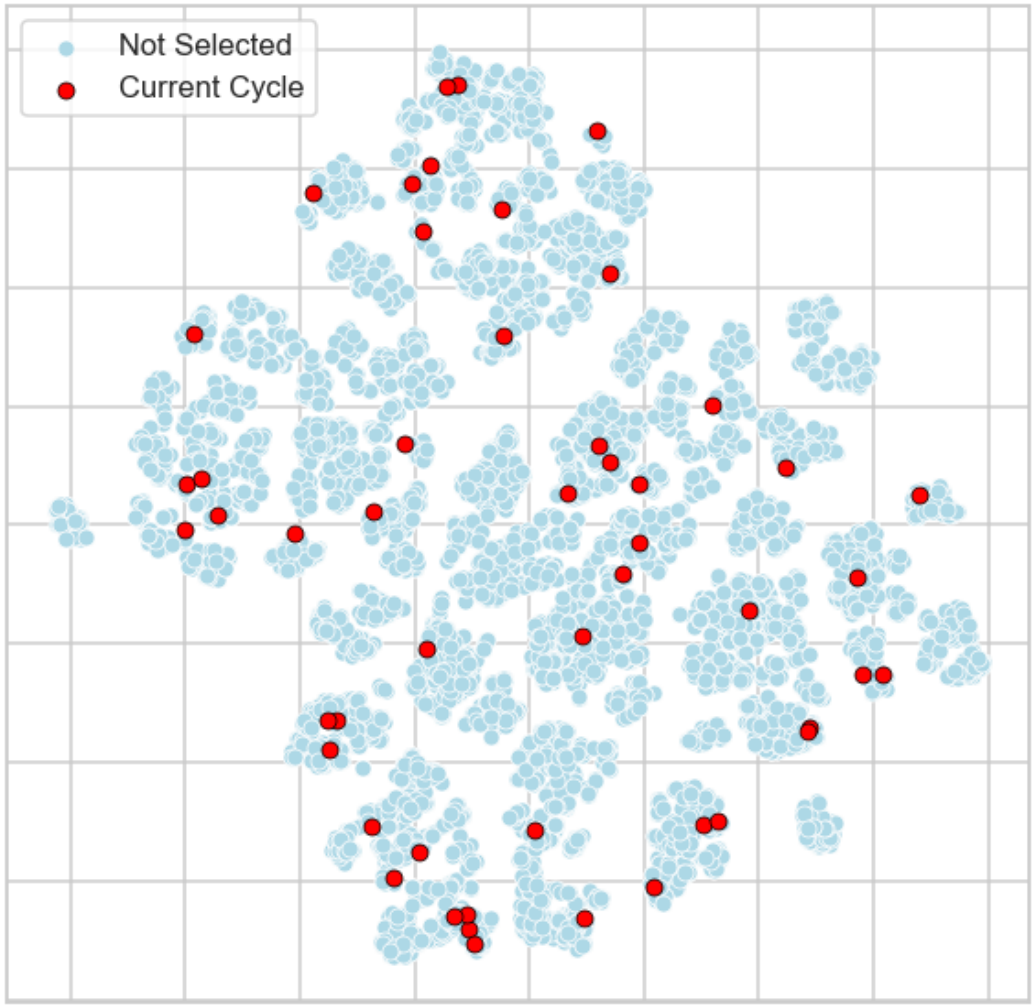
ΔG = f(molecular descriptors)

Predict ΔG for the remaining 9,950 molecules

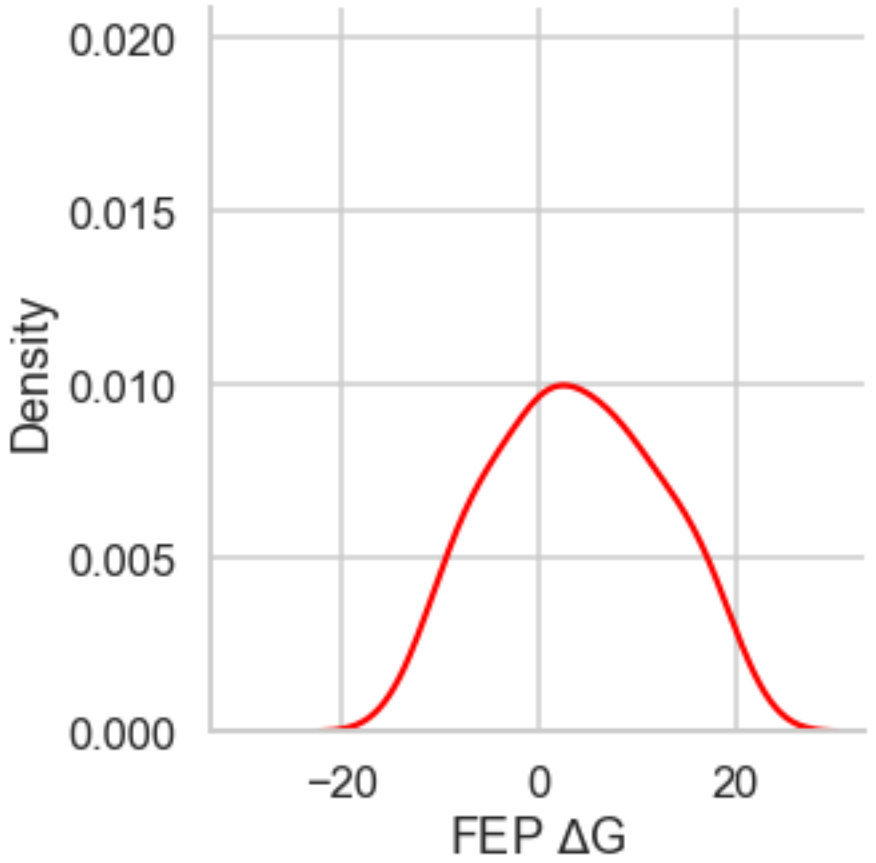
Run FEP on the 50 molecules with the best predicted ΔG

Repeat

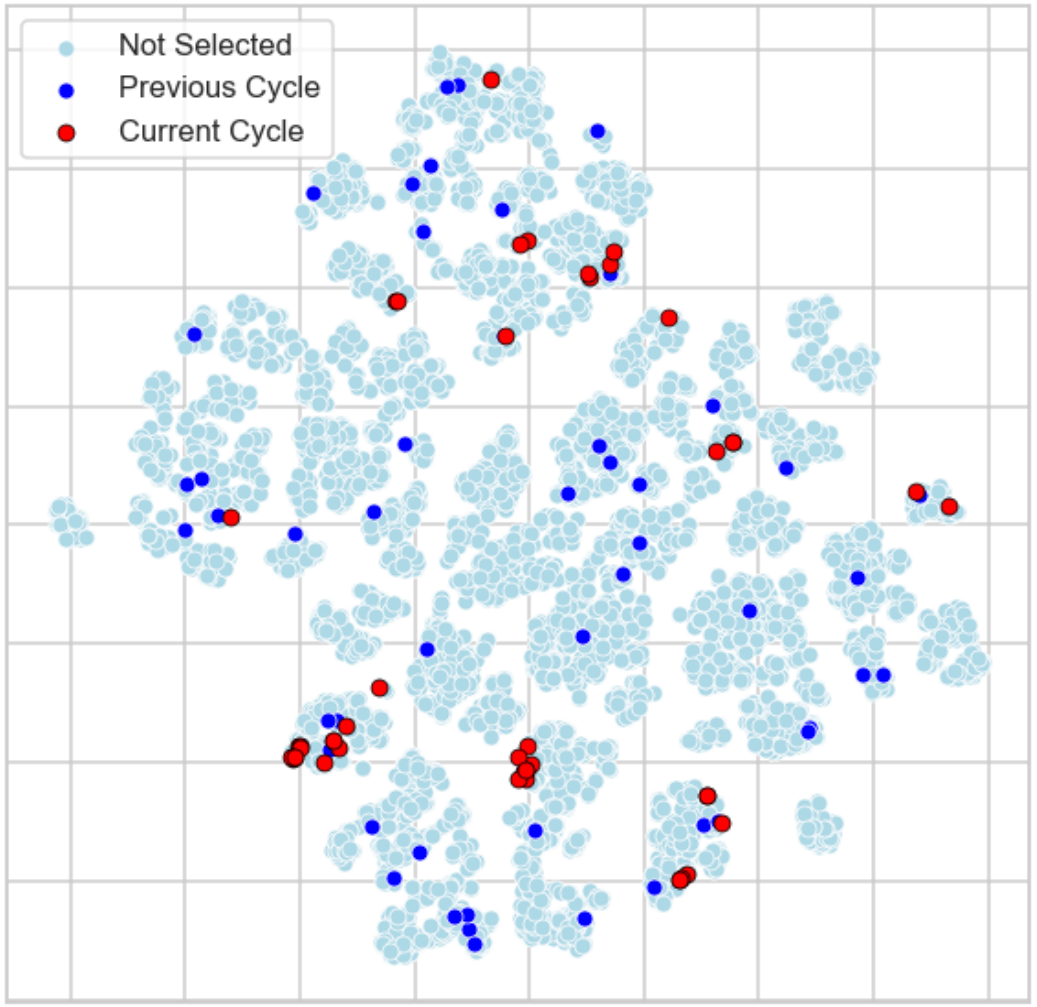
Initial Sample - 1 of the Top 100 Molecules Found



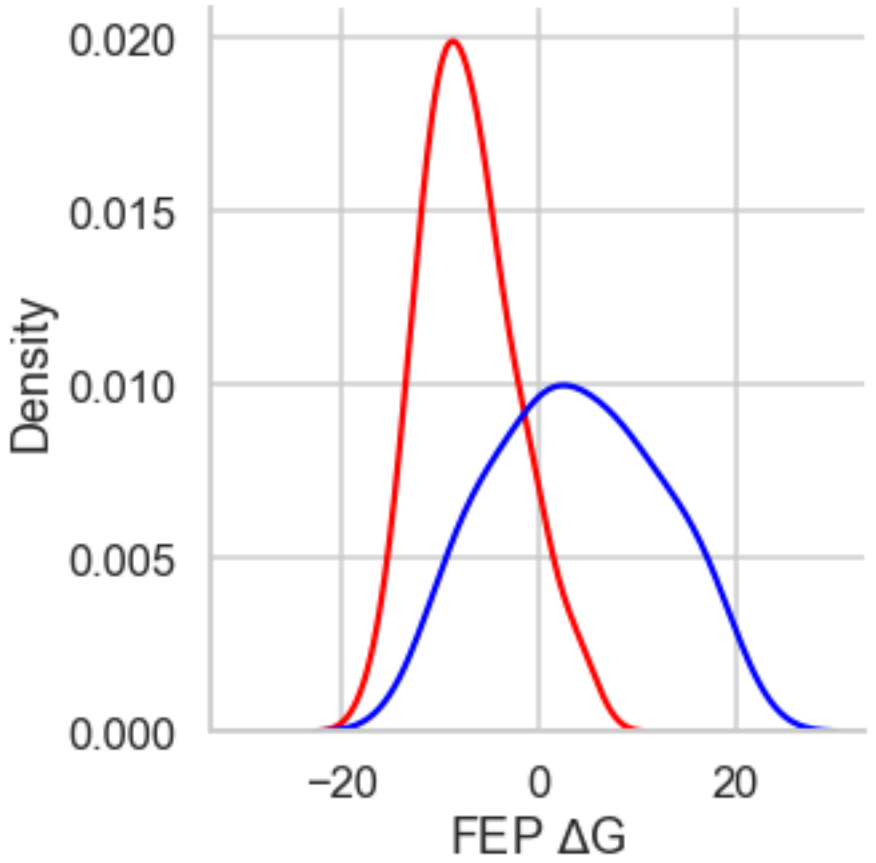
Sampled 0.5%



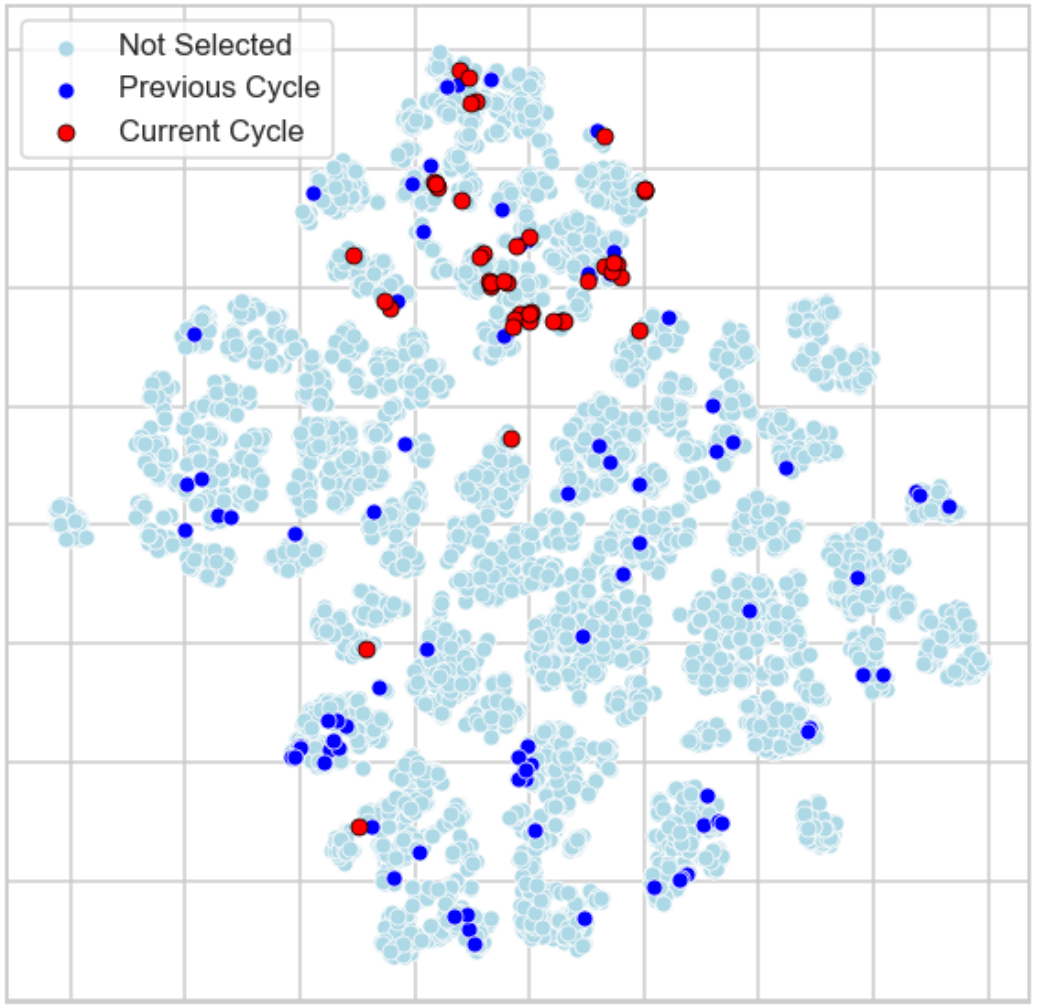
Cycle 1 - 16 of the Top 100 Molecules Found



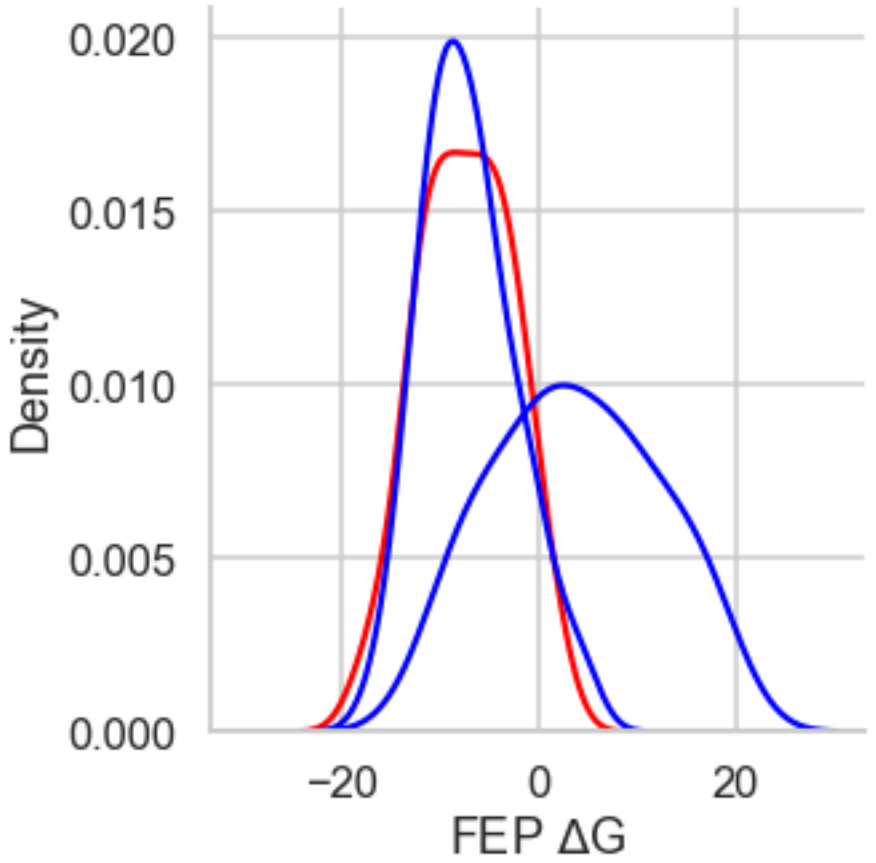
Sampled 1%



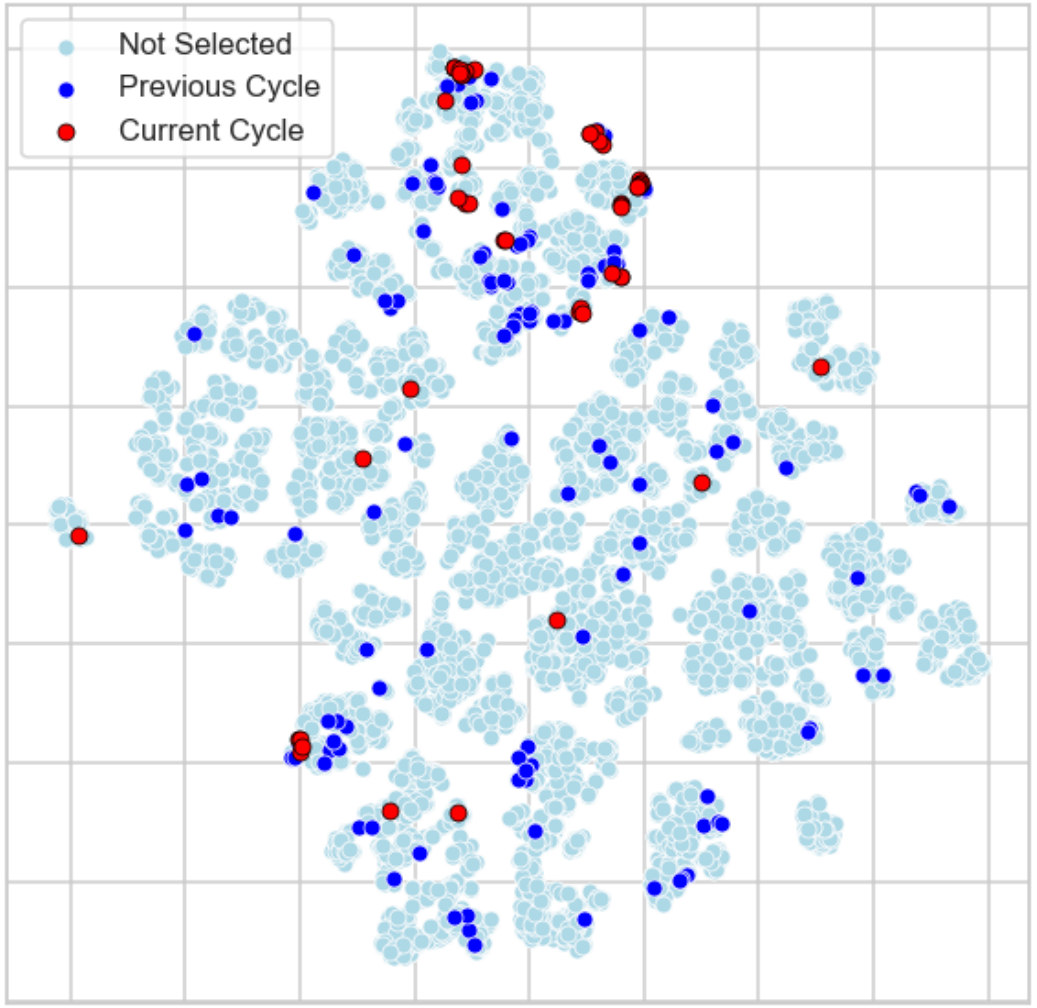
Cycle 2 - 31 of the Top 100 Molecules Found



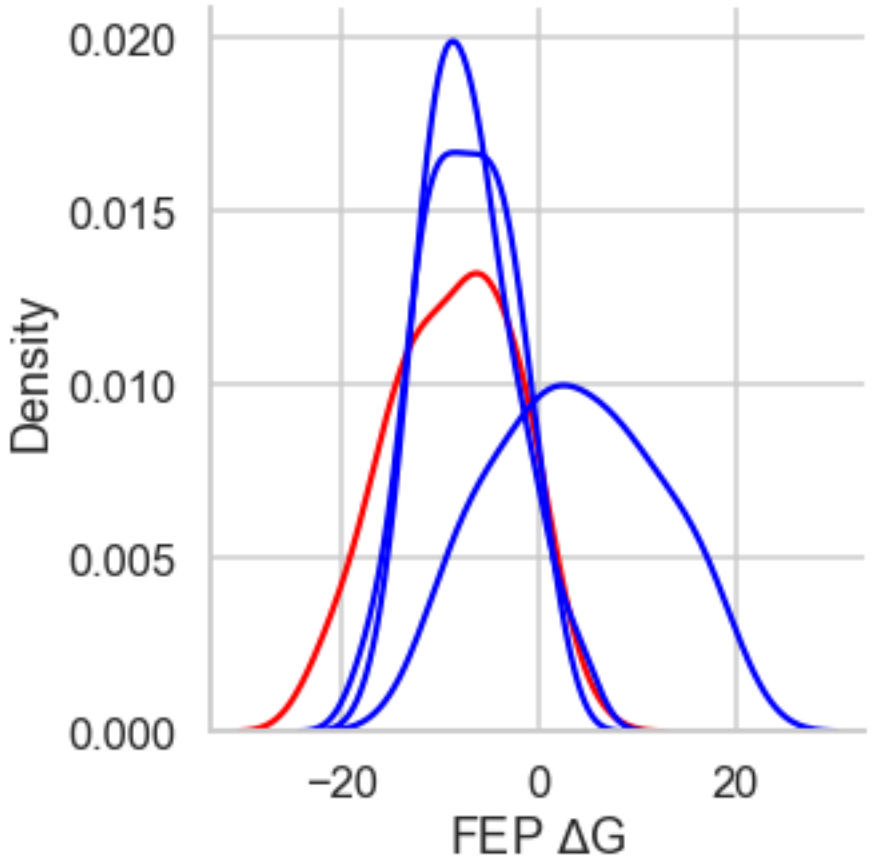
Sampled 1.5%



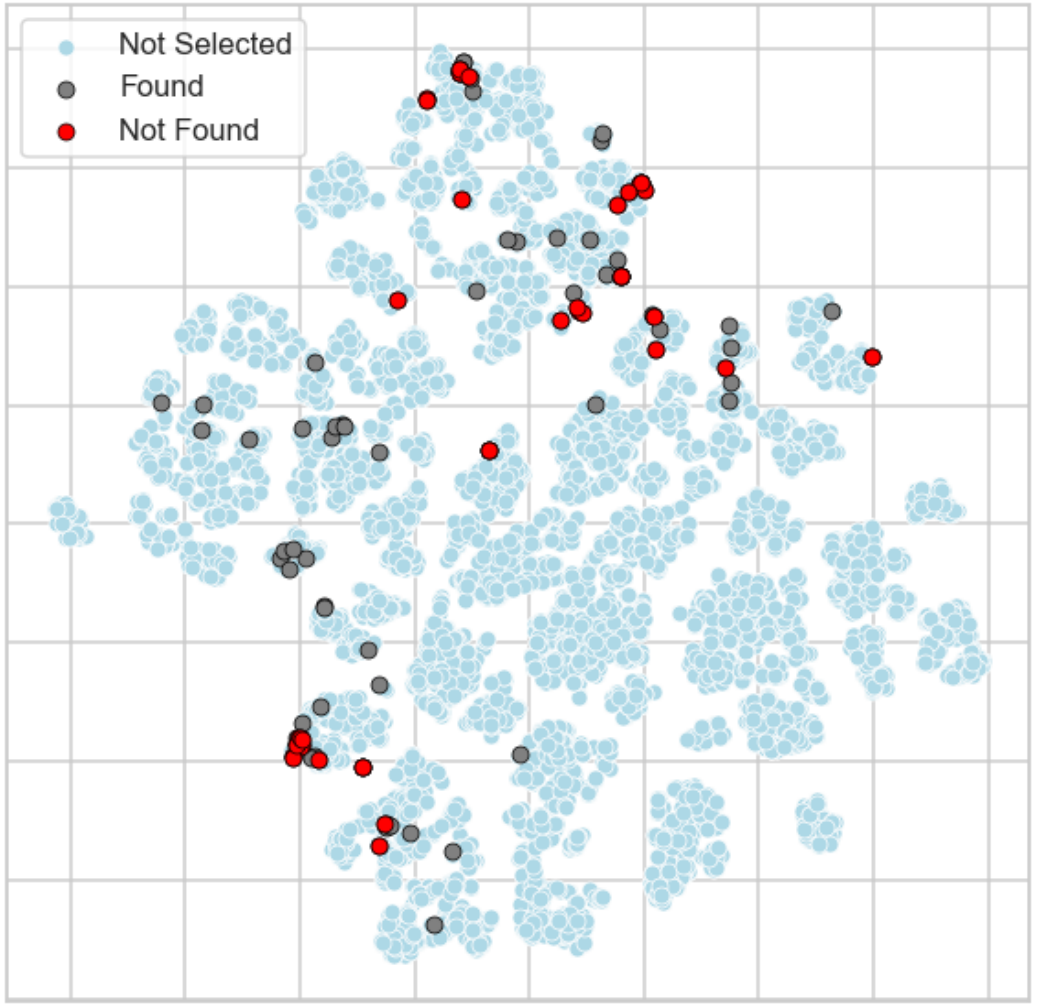
Cycle 3 - 37 of the Top 100 Molecules Found

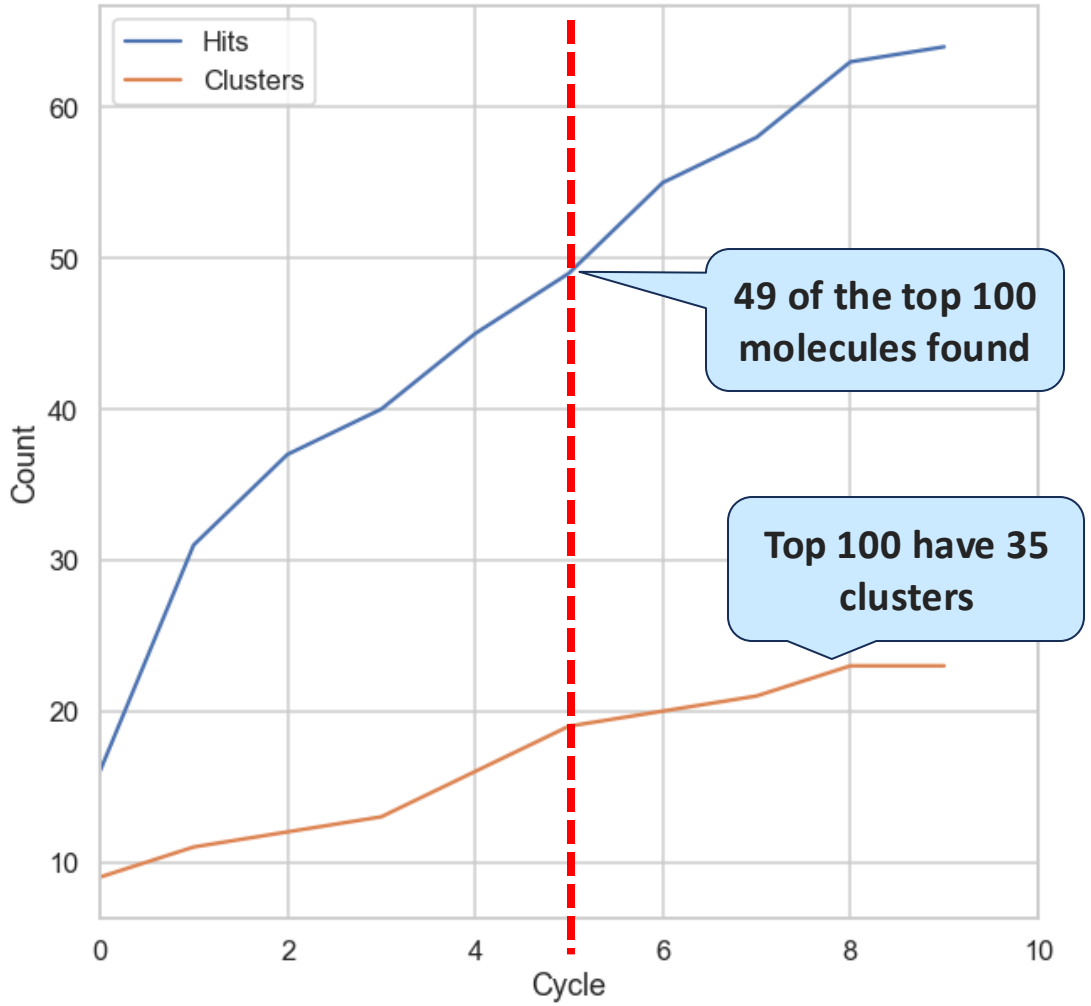


Sampled 2%

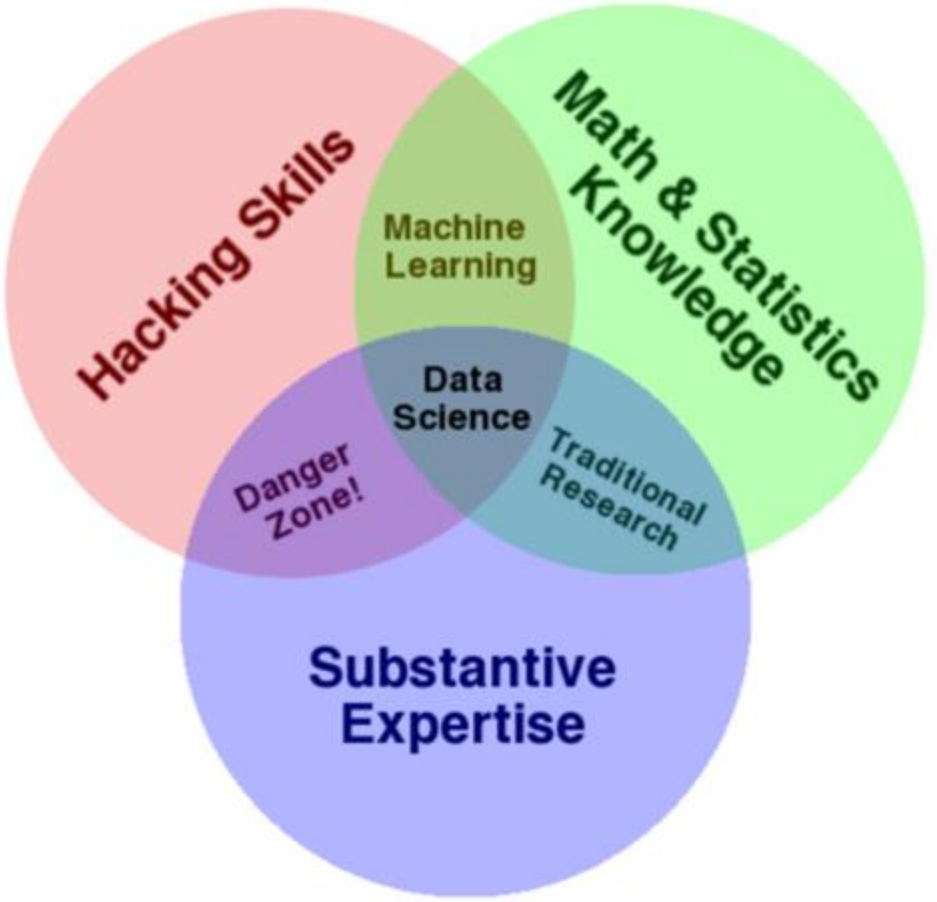


Comparing Top 100 Compounds Found and Not Found in 300 Samples (3%)



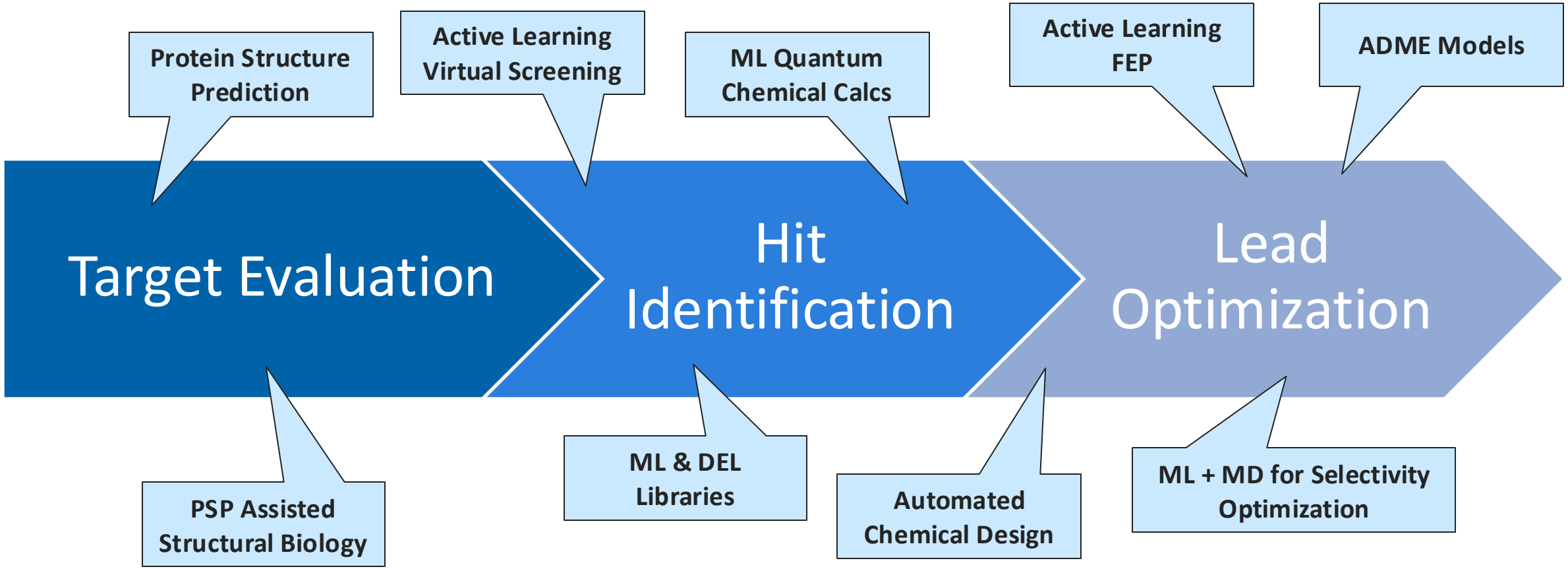


What Do We Need To Succeed?



http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram

ML Has Impact Across the Drug Discovery Process



ML is One Component in a Collaboration Between Experiment and Computation