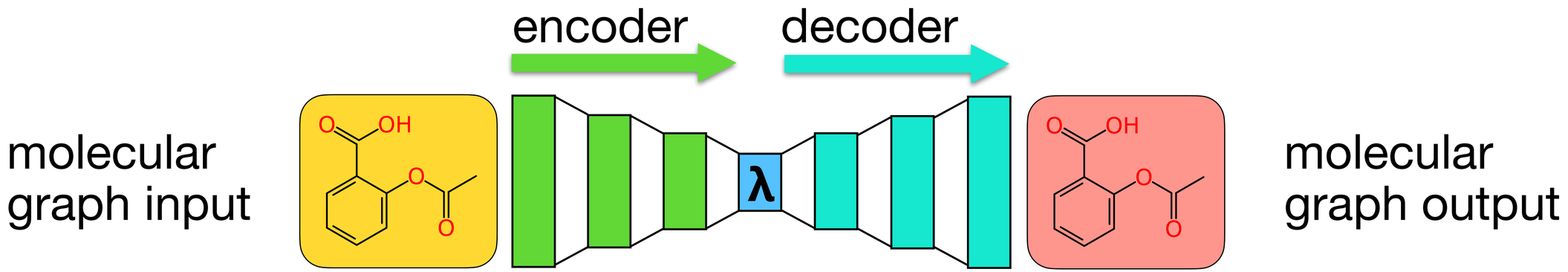
Some of the doors we unlock by describing molecules using generative models

Alex M. Clark

alex@collaborativedrug.com

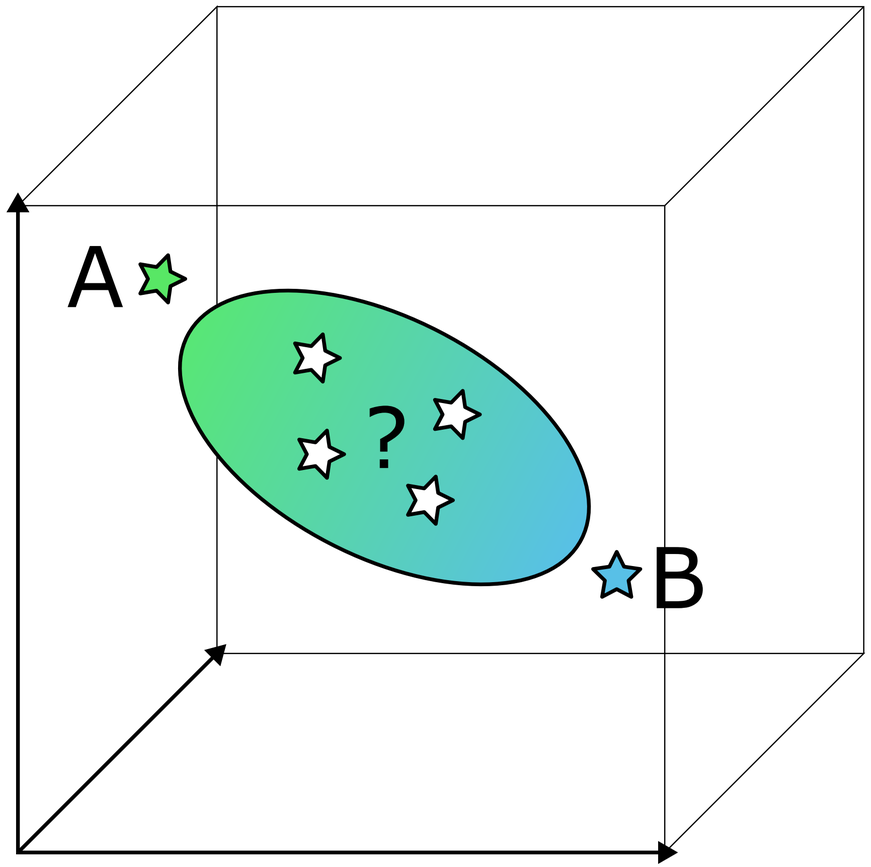
Generative Models



- Bottleneck is the *latent vector* (**λ**):
  - highly orthogonalized fixed-length vector descriptor
  - *and* any set of values can be used to recreate corresponding molecule
- Most models use SMILES input & output
  - linear representation is a good fit for the technology
  - ours also uses graph inputs and fingerprint outputs

Navigation of Chemical Space

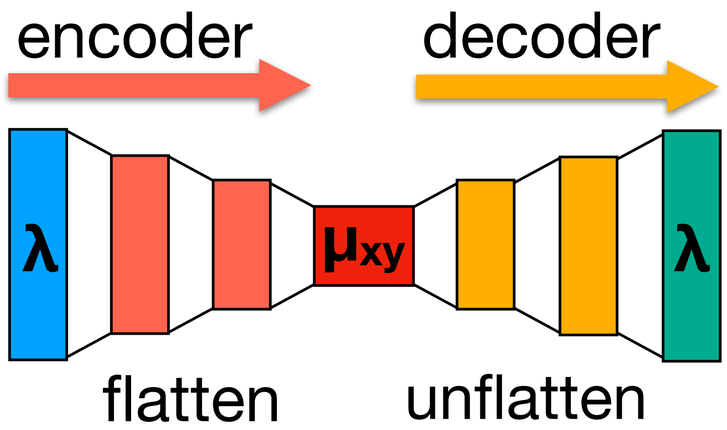
- **λ** as a multi-dimensional coordinate



- e.g. if molecule **A** has good activity and **B** has good ADME, is there something good in between?

Too Many Dimensions

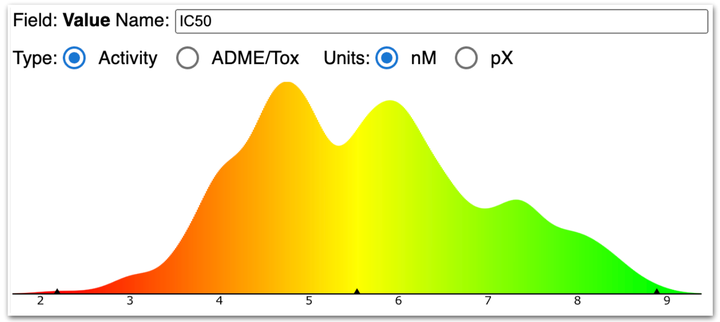
- λ_{384} necessary to get most of ChEMBL ⇛ SMILES & ECFP
- Many dimensions hard to visualize: reduction from 384 → 2

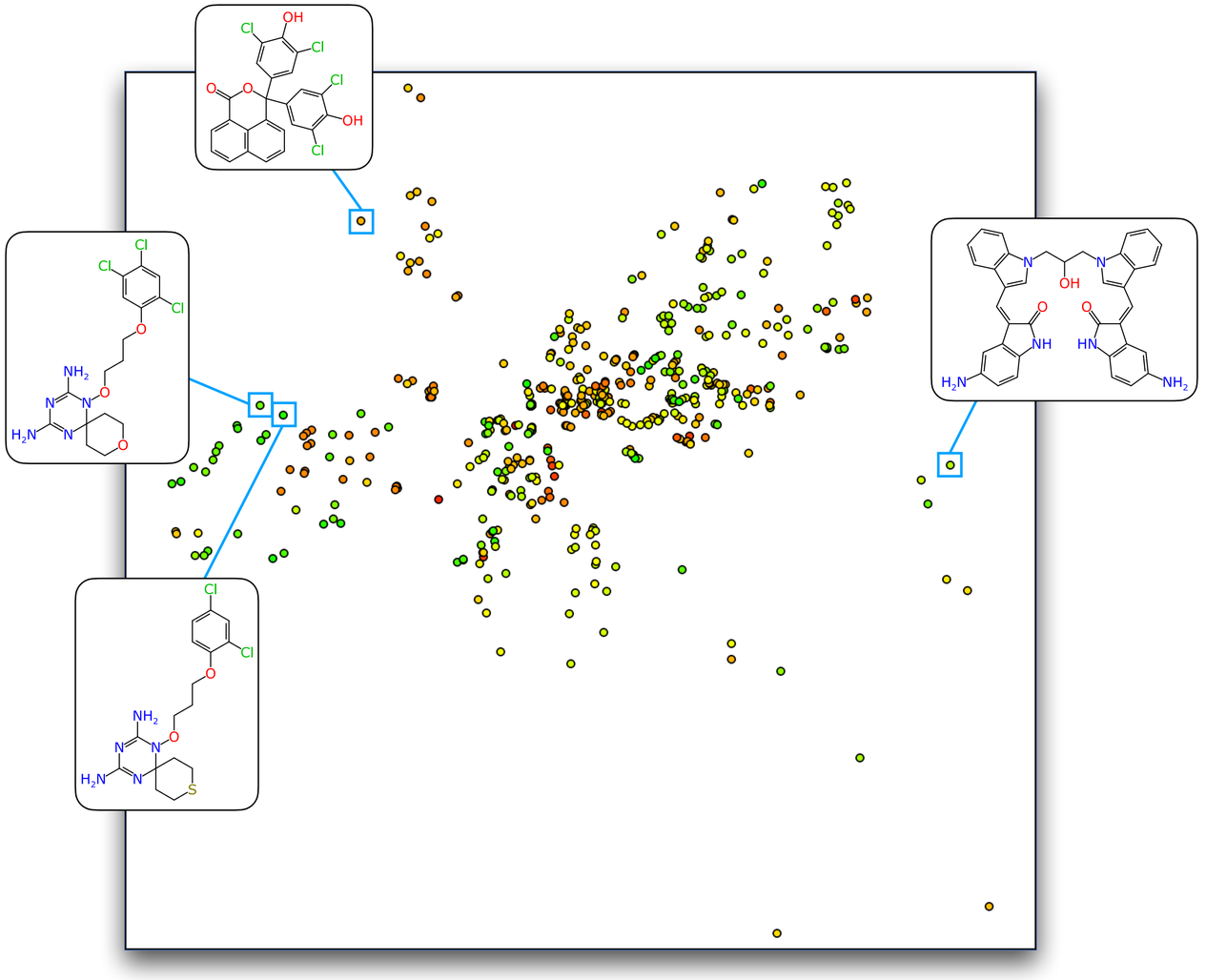


- Medium sized datasets (several thousand molecules) work well
- Very simple neural network construction
- Transformation to μ_{xy} is ***reversible***

Flattening Model

DHFR, 500 molecules

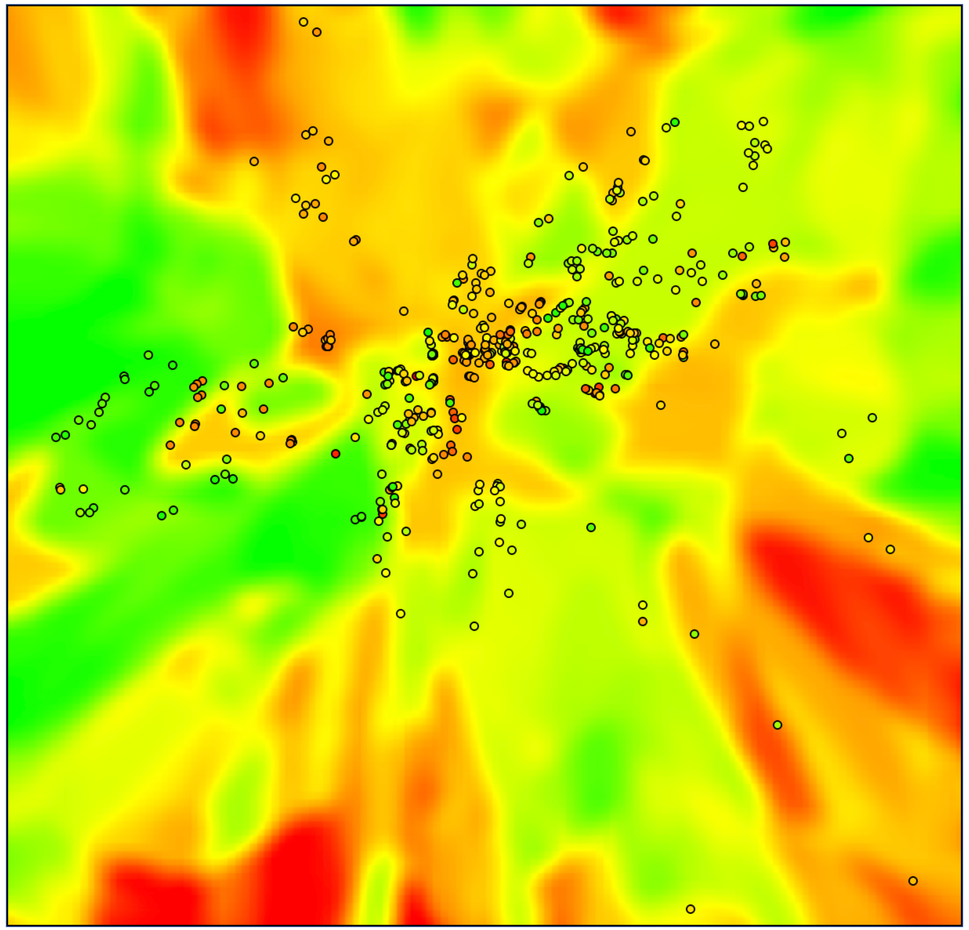




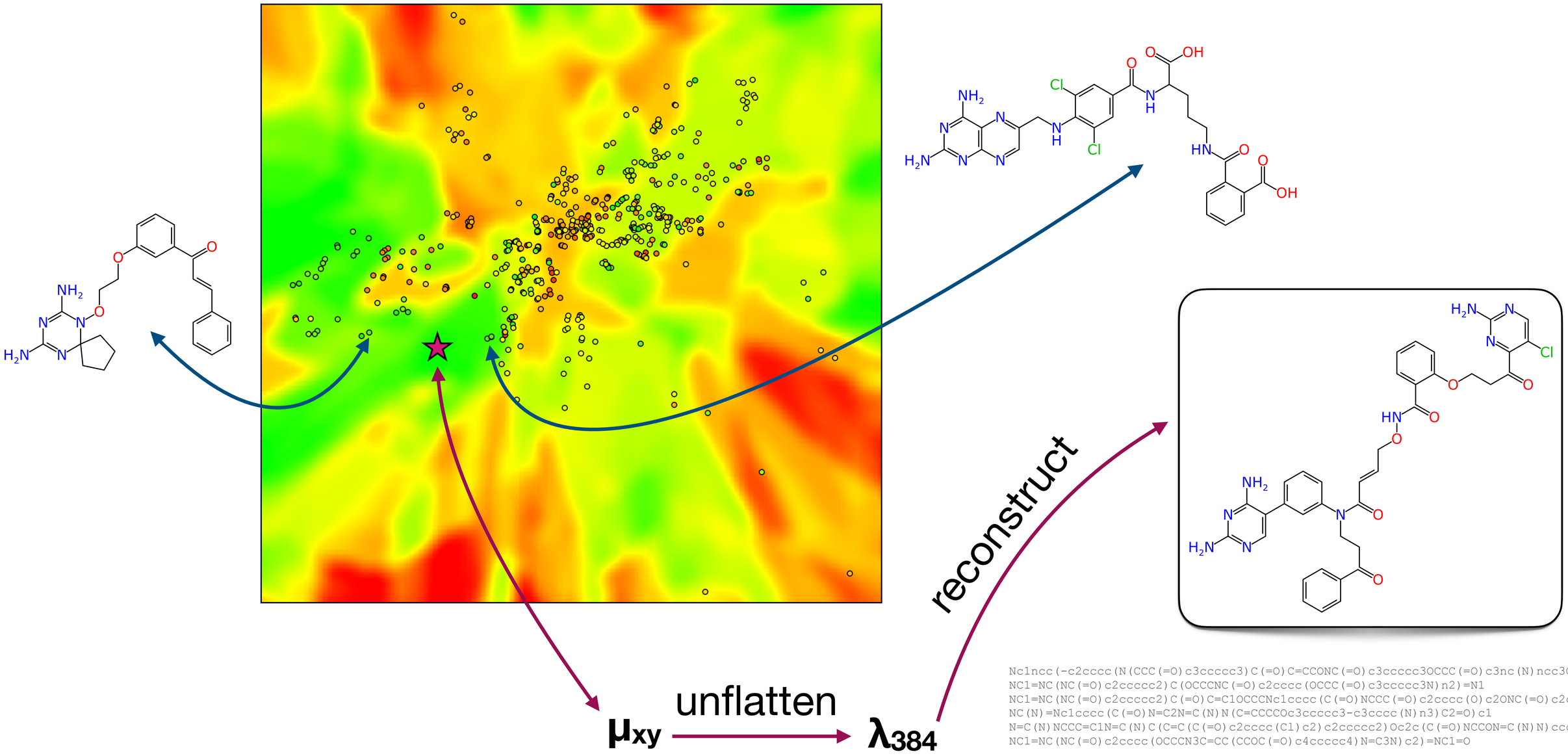
- λ_{384} vectors reversibly projected onto 2D, displayed interactively
- Fun fact: deep learning model is *trained* in-browser, native JavaScript

Activity Contours

- Create a "gradient boost" model for $\boldsymbol{\lambda_{384}} \rightarrow$ **activity** (fast to train)
- Empty grid: *unflatten* $\boldsymbol{\mu_{xy}} \rightarrow \boldsymbol{\lambda_{384}}$ and feed into model... colour the grid

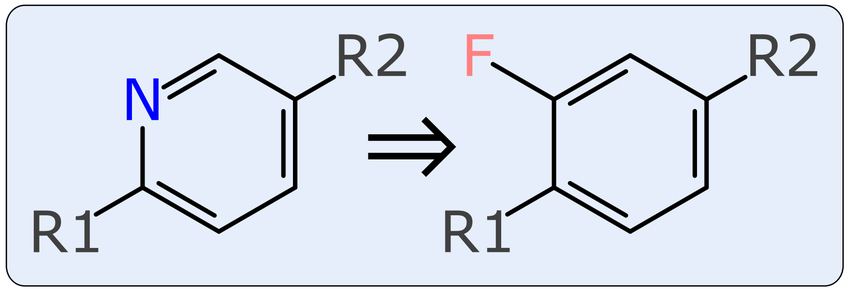


Exploring




```
Nc1ncc(-c2cccc(N(CCC(=O)c3ccccc3)C(=O)C=CCONC(=O)c3ccccc3OCCC(=O)c3nc(N)ncc3Cl)c2)c(N)n1
NC1=NC(NC(=O)c2ccccc2)C(OCCCNC(=O)c2cccc(OCCC(=O)c3ccccc3N)n2)=N1
NC1=NC(NC(=O)c2ccccc2)C(=O)C=C1OCCCNc1cccc(C(=O)NCCC(=O)c2cccc(O)c2ONC(=O)c2ccccc2)n1
NC(N)=Nc1cccc(C(=O)N=C2N=C(N)N(C=CCCCOc3ccccc3-c3cccc(N)n3)C2=O)c1
N=C(N)NCCC=C1N=C(N)C(C=C(C(=O)c2cccc(Cl)c2)c2ccccc2)Oc2c(C(=O)NCCON=C(N)N)cccc21
NC1=NC(NC(=O)c2cccc(OCCCN3C=CC(CCOC(=O)c4ccccc4)N=C3N)c2)=NC1=O
```

Bio-isosteres

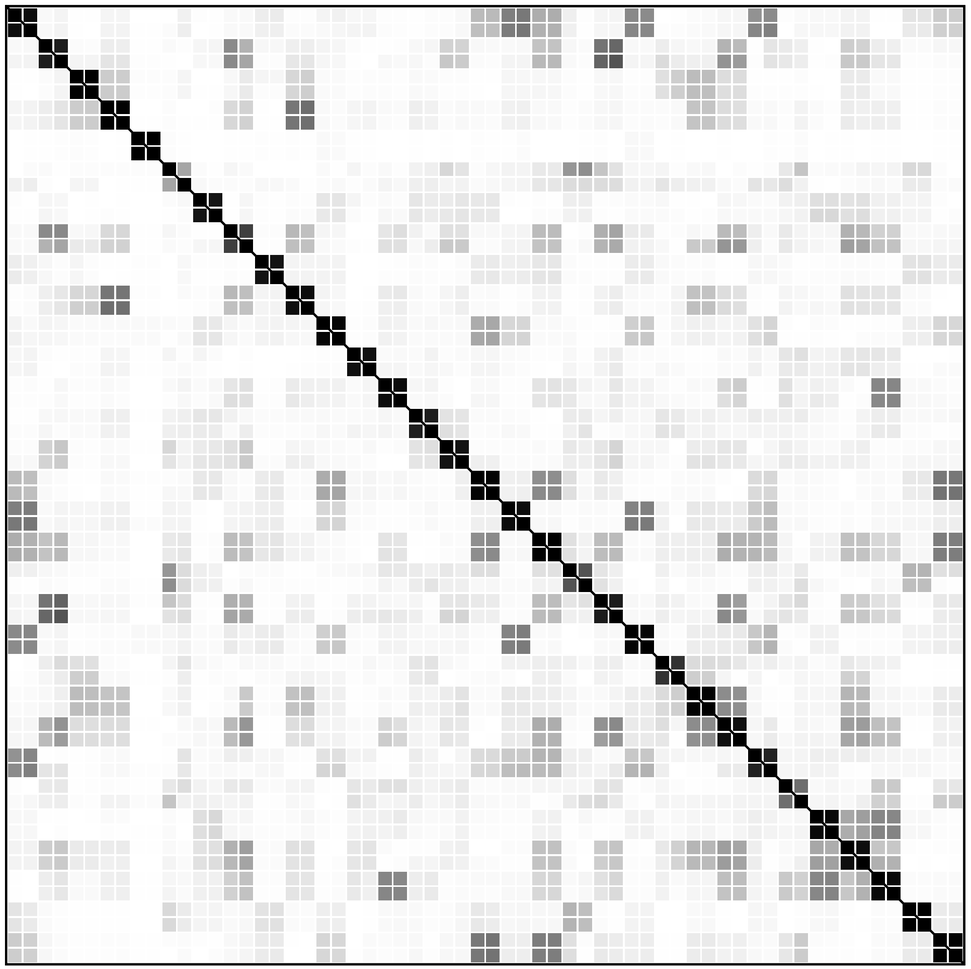


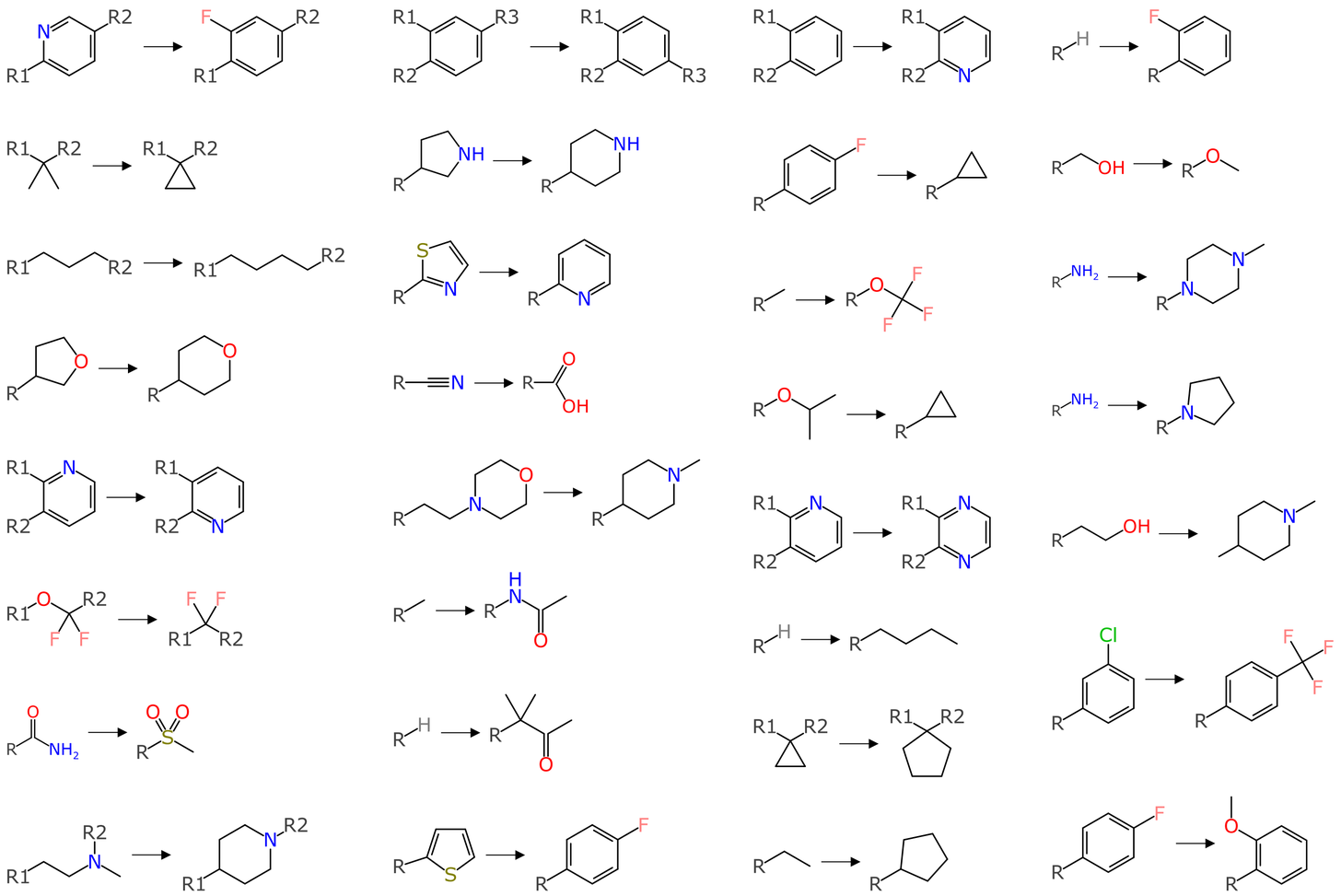
- Modify the structure to:
  - keep the biological activity
  - re-roll the dice on everything else
- Conventionally implemented as scaffold replacements

- Pain points:
  - where to get suitable bio-isostere transforms?
  - what if you need something a bit less cleanly defined?

Quasi-bioisosteres?

Orthogonality

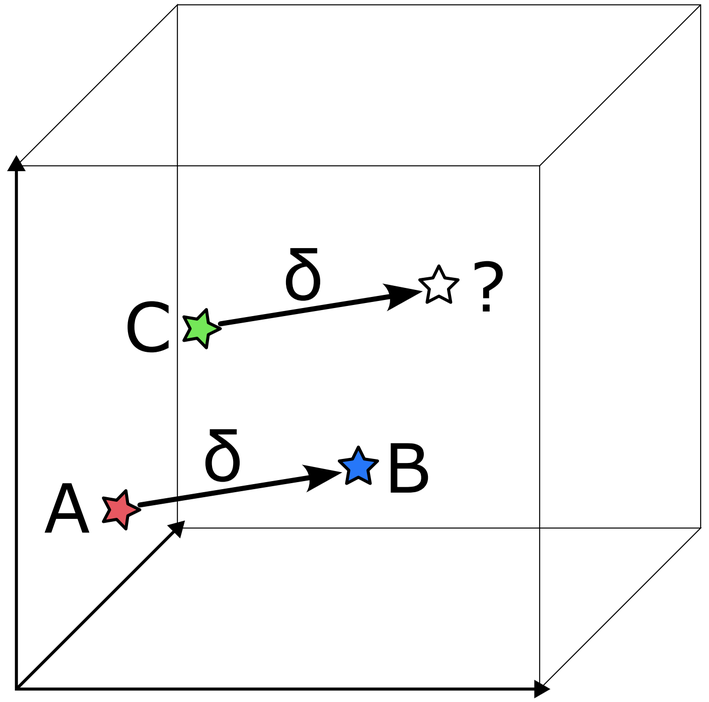




indication of how similar transforms are to each other, in latent space

Why?

- Propose a *structural transformation*:
  - has bioisostere-like properties
  - but not necessarily a well defined substructure swap



- Molecules **A** and **B** have good activity, but **B** has better ADME
- Define transform **δ** as **λ(A) ⇒ λ(B)** in latent vector space
- Apply transform as **λ(C) + δ** to improve ADME...?

passes the eyeball test, but needs to be developed

Conclusions & Future Work

- Operating in *latent vector space* enables new drug design tools...
- ... currently in early prototype stage
- Generative models: shrink latent vectors? alternatives to SMILES?
- Exploring multidimensional space via gradient
- Contours: plot multiple activities and ADME/tox
- Quasi-bioisosteres: can it enrich datasets?
- TODO: get some of these tools into the hands of scientists...

Questions?

- Contact:
  - Alex M. Clark **alex@collaborativedrug.com** (Collaborative Drug Discovery)
- Thanks to the Vault & Research Informatics teams