Mining Novel Biomaterials

with scikit-learn

Materials Informatics

- Think bioinformatics, but using materials data.
- http://www.materialsproject.org/: open database of calculated materials data

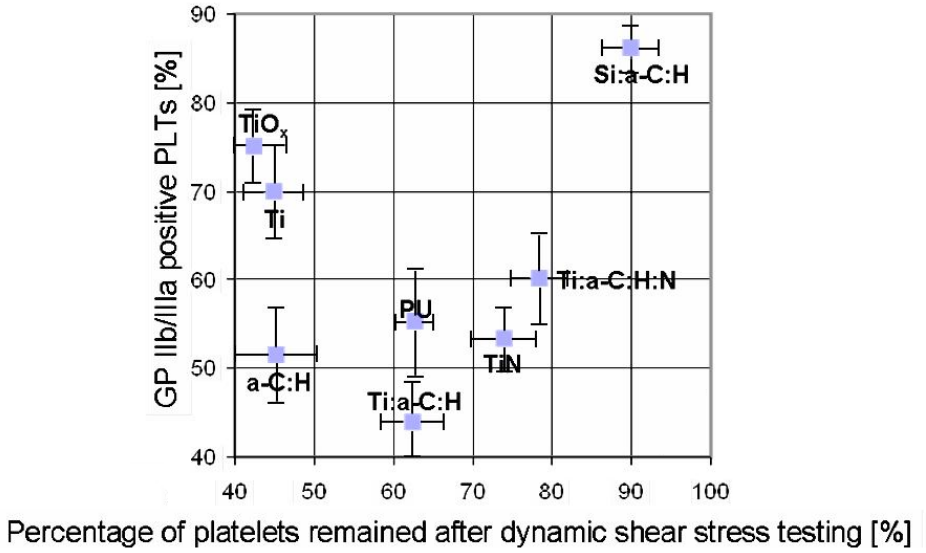
- Could we do something interesting with it?

Biocompatibility

- Does a material cause harm when introduced in the human body?
- Ill-defined; a more specific goal is hemocompatibility.
- How does a material interact with blood platelets?

Training Data: Labels

- Each dot is a material
- Upper left = bad
- Lower right = good
- Score = x - y



Training data: Features

- From the Materials Project API
- Four material parameters:
  - band_gap
  - e_above_hull
  - energy_per_atom
  - formation_energy_per_atom
- Standardisation with scikit-learn

Model

- Linear regression
- Best materials have same elements as training set
- Si seems worth exploring
- TODO: more data, implement CV

Chemical Formula	Model Score
Si_2CN_4	0.385
$Si(CN_2)_2$	0.383
Si_3N_4	0.347
Si_2N_2O	0.298
P_4N_6O	0.262
V_6C_5	0.256
V_2C	0.253
PNO	0.246
V_8C_7	0.243