COMP 150 CSB –
Computational Systems Biology

Analysis of Enzyme Data & Enzyme Promiscuity

Soha Hassoun

Department of Computer Science (primary)

Department of Chemical and biological Engineering

Department of Electrical and Computer Engineering

Tufts UNIVERSITY

Outline

- Where to find papers?
- Enzyme Kinetics: quick review
- Data analysis on kinetics for natural substrates,

  Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., & Milo, R. (2011). The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry, 50(21), 4402-4410.

  - Goals of the study
    - Global analysis of kinetic parameters based on very large set of data from BRENDA
    - Identify data trends and correlations, pose some hypothesis
  - Selected because it provides nice analysis
- What about enzymes acting on non-natural substrates?
  - Enzyme promiscuity
- Homework #2

Where to get papers and supplementary material?

- scholar.google.com
  - Clicking on links -- Access to many of the articles when connected through campus
- Jumbo Search
  - Yes, it is actually good, with links
- PubMed – the main search engine
  - https://www.ncbi.nlm.nih.gov/pubmed/
  - Find article by name
- Where to find Supplementary material/info?
  - Through links on publication website
  - Or by hyperlnks
  - Search for Supporting info, or Supplementary
  - Example: https://pubs.acs.org/doi/abs/10.1021/bi2002289

Enzyme Kinetics - Clarification

- Write the rate of change of S:

  $d[S]/dt = - k_1[E][S] + k_{-1}[ES]$

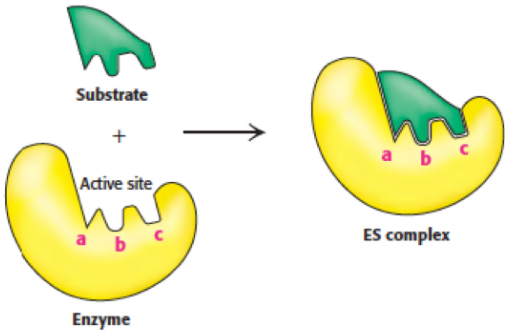
- What are units of k_1 and k_{-1} ?
- k_1 has units of concentration^{-1}time^{-1}
- k_{-1} has units of time^{-1}

- Similarly, k_2 has units of time^{-1}

- $K_M = (k_{-1} + k_2) / k_1$ will have units of concentration

Enzyme Kinetics - refresher

- Turnover number, k_{cat}, same as k_2
- Michaelis constant: $K_M = (k_{-1} + k_2) / k_1$
  - When k_2 is small, $K_M \cong k_{-1} / k_1 = K_s$ (the enzyme-substrate dissociation constant)
- $k_{cat}/K_M = (k_1 * k_2) / (k_{-1} + k_2)$
  - If k2 is small, then k_{cat}/K_M is small
  - If K2 is large, then k_{cat}/K_M is limited by the value of k1 (or the rate of formation of ES, and rate of diffusion of S)



$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} ES \xrightarrow{k_2} E + P$$

Enzyme Kinetics - refresher

- Diffusion limits the value of k_1, the rate of formation of ES
  - It cannot be higher than 10^8 and 10^9 s^{-1} M^{-1}

$$k_{cat}/K_M = \frac{k_{cat}k_1}{k_{-1} + k_{cat}} = \left(\frac{k_{cat}}{k_{-1} + k_{cat}}\right)k_1$$

  - So k_{cat} / K_M can at most be 10^8 and 10^9 M^{-1}

Data for the study

- BRENDA 2010 data
- Enzymatic data for natural and non-natural substrates
  - Use KEGG 2010 data to identify natural reactants
  - Removed co-factors because:
    1. They participate in multiple enzymatic reactions
    2. Co-factors are subject to a different evolutionary pressures compared to other substrates
- Around 90,000 entries for K_M and 35,000 entries for k_{cat} (Table 1).

Data is noisy

- k_{cat} and K_M values from different studies varies by a factor of 2.5-2.9
- Solution: take the median
- Other issue:
  - Correlation between the k_{cat} values for different substrates of the same multi-substrate reaction (expected to be identical) are not correlated (R^2 = 0.63)
  - Why? See next slide

Reasons for Noisy Data

- Measurements under different conditions (pH, Temp, ionic strength, metal-ion and co-factor concentrations)
- Differences between *in vitro* and *in vivo* parameters (as much as 3 orders of magnitude)
- 20% of the values in the Brenda database do not correspond to the values reported in the corresponding reference papers
- While analysis is global, there are subgroups that do not display identified correlation
  - Example: complex mechanisms do not correlate between molecular weight, K_M,K_{cat}. Inclusion lowers observed correlations

The Average Enzyme Is Far From Kinetic Perfection

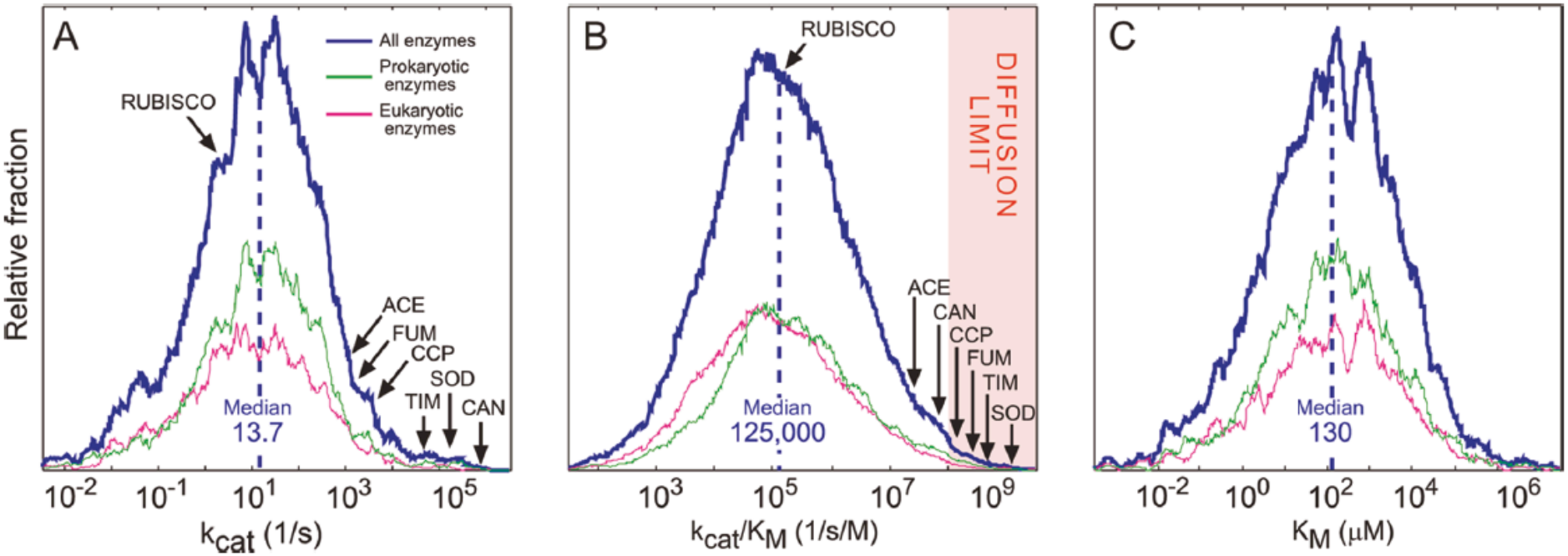
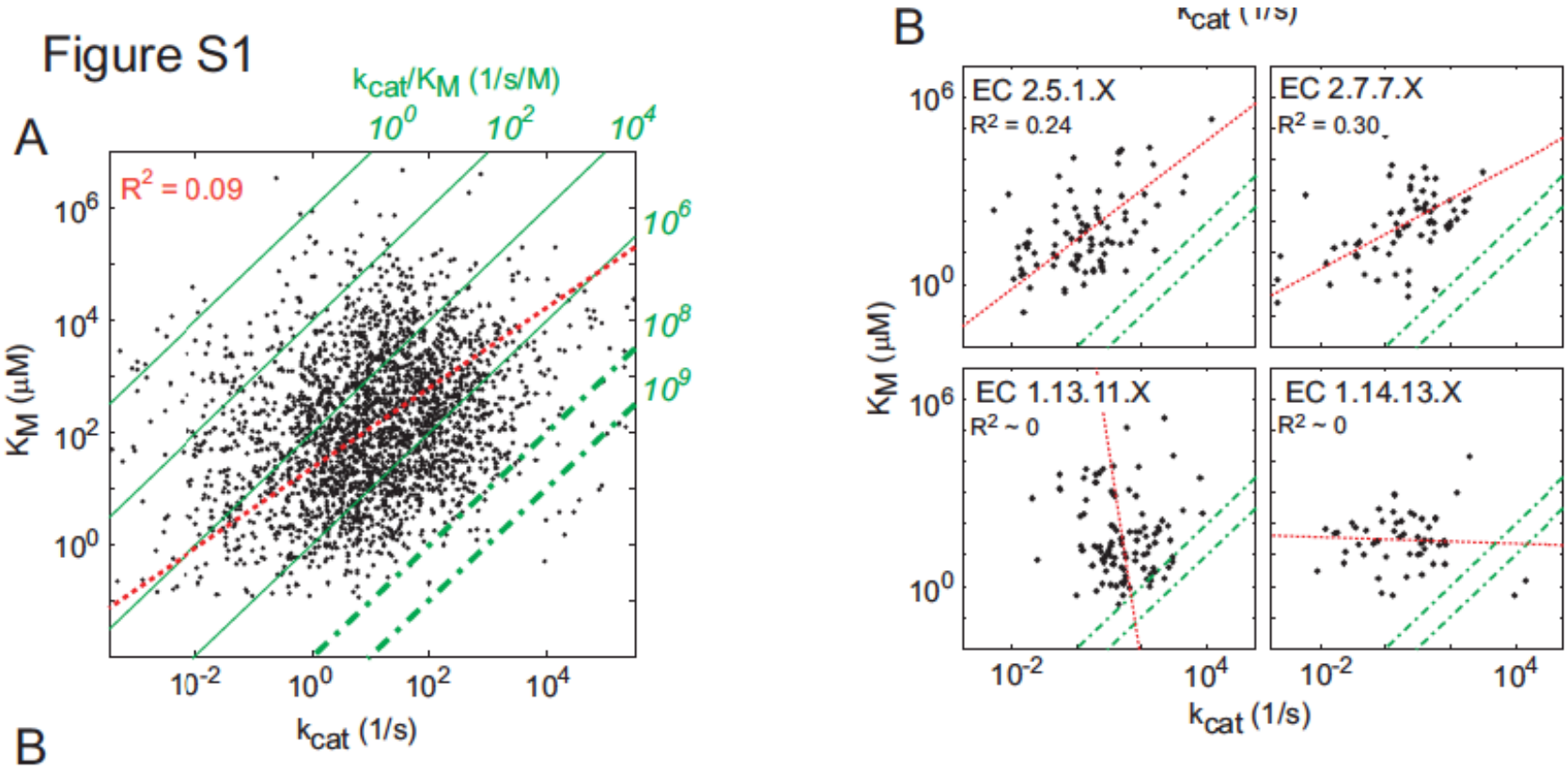


Figure 1. Distributions of kinetic parameters: (A) k_{cat} values ($N = 1942$), (B) k_{cat}/K_M values ($N = 1882$), and (C) K_M values ($N = 5194$). Only values referring to natural substrates were included in the distributions (Supporting Information). Green and magenta lines correspond to the distributions of the kinetic values of prokaryotic and eukaryotic enzymes, respectively. The location of several well-studied enzymes is highlighted: ACE, acetylcholine esterase; CAN, carbonic anhydrase; CCP, cytochrome *c* peroxidase; FUM, fumarase; Rubisco, ribulose-1,5-bisphosphate carboxylase oxygenase; SOD, superoxide dismutase; TIM, triosephosphate isomerase.

Bar-Even, A., Noor, E., Savir, Y., Liebermeister, W., Davidi, D., Tawfik, D. S., & Milo, R. (2011). The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters. Biochemistry, 50(21), 4402-4410.

Dependency between K_{cat} and K_m



- Each dot is a enzyme-substrate-organism
- Dotted green are diffusion limit range
- Red lines are trend lines, calculated orthogonal least square fitting
- Weak positive correlation (0.09)

The correlation between k_{cat} and K_M (A) across all reactions (R^2=0.09) and (B) for reactions associated with specific EC classes. Green corresponds to k_{cat}/K_M isovalue lines. Bold, dashed lines represent the diffusion limit for small metabolites interacting with proteins ($k_{cat}/KM<10^8$-10^9). Red corresponds to the trend lines, calculated by orthogonal least-square fitting. Each dot represents a combination of an enzyme, a substrate and an organism. EC classes: 2.5.1.X (R^2=0.24): Transferase enzymes, transferring alkyl or aryl groups, other than methyl groups; 2.7.7.X (R^2=0.30): Nucleotidyltransferases; 1.13.11.X (R^2~0): Oxidase enzymes, acting on a single donor and incorporating two atoms of oxygen and 1.14.13.X (R^2~0): Oxidase enzymes, acting on two donors, where NADH or NADPH is one donor,

Selective pressures mold enzyme kinetics

- Key question:

Why do most enzymes operate far from the theoretical limit?

- Hypothesis:
  - Due to evolutionary pressures
  - Maximal rates have not evolved in cases where a particular enzyme stringent selection

- Evaluation:
  - Classified modules into four primary groups:

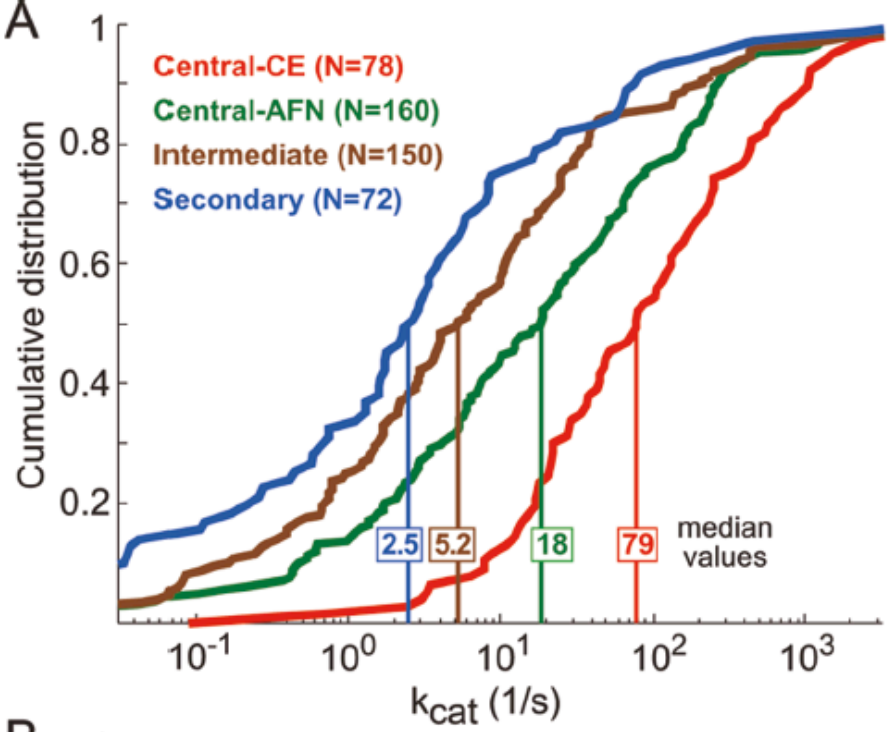
Table 2. Classification of Several Example Metabolic Pathways[a]

metabolic groups	example of pathways
primary metabolism—CE (carbohydrate and energy)	glycolysis/gluconeogenesis
	pentose phosphate pathway
	citrate cycle (TCA cycle)
	carbon fixation in photosynthetic organisms
	pyruvate metabolism
	glyoxylate and dicarboxylate metabolism
	etc.
primary metabolism—AFN (amino acids, fatty acids, and nucleotides)	glycine, serine, and threonine biosynthesis and metabolism
	phenylalanine, tyrosine, and tryptophan biosynthesis and metabolism
	cysteine and methionine biosynthesis and metabolism
	fatty acid biosynthesis
	purine metabolism
	pyrimidine metabolism
	etc.
intermediate metabolism	pantothenate and CoA biosynthesis
	ubiquinone and other quinone biosynthesis
	glutathione metabolism
	biotin metabolism
	folate biosynthesis
	thiamine metabolism
	etc.
secondary metabolism	flavonoid biosynthesis
	taurine and hypotaurine metabolism
	phenylpropanoid biosynthesis
	bile acid biosynthesis
	caffeine metabolism
	retinol metabolism
	etc.

[a] The full list, containing the classification of 300 metabolic modules, is given in Table 2 of the Supporting Information.

Selective pressures mold enzyme kinetics

- K_{cat} for central-CE is 30x higher than that with secondary metabolism
- Central-CE enzymes are ~6x higher than intermediate and secondary ones



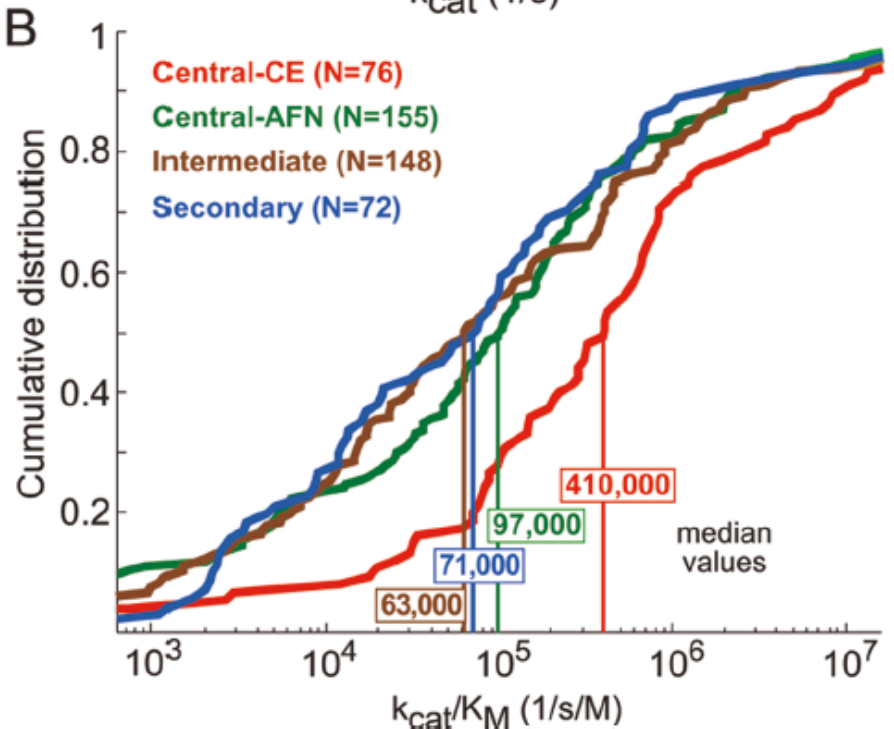


Figure 2. Enzymes operating within different metabolic groups have significantly different k_{cat} and k_{cat}/K_M values. (A) Distribution of k_{cat} values for enzyme–substrate pairs belonging to different metabolic contexts. All distributions are significantly different with a p value of <0.0005 (rank-sum test), except for intermediate versus secondary metabolisms ($p < 0.05$). (B) Distribution of k_{cat}/K_M values for enzyme–substrate pairs belonging to different metabolic contexts. Central-CE (carbohydrate and energy) metabolism has significantly higher k_{cat}/K_M values than all other metabolic groups [$p < 0.0005$ (rank-sum test)]. Abbreviations: CE, carbon and energy; AFN, amino acids, fatty acids, and nucleotides. Numbers in parentheses represent the numbers of enzyme–substrate pairs included in each set.

Selective pressures mold enzyme kinetics

- Central metabolism requires high average flux rates to enforce molecular production
  - Required improved kinetic parameters that may reduce the amount of enzymes needed and hence alleviate the cost

- The selection pressure might be weaker in enzymes functioning in secondary metabolism
  - Optimality of secondary metabolism is not a measure of organism fitness
  - Secondary metabolism operates under specific conditions and often for short periods of time and with lower fluxes
  - Perhaps secondary metabolism evolved for regulation, control, etc.
  - Data is noisy – perhaps the natural products in BRENDA for secondary metabolism enzymes are not correct

- The differences in k_{cat} values between the metabolic groups are more prominent than the differences in the k_{cat}/K_M value:
  - Because of demands for higher fluxes, k_{cat} is important.
  - An enzyme, under a constant k_{cat} / K_M , is expected to increase its kcat at the expense of K_M .

Examine relationship between K_{cat} and EC classes

Figure S5

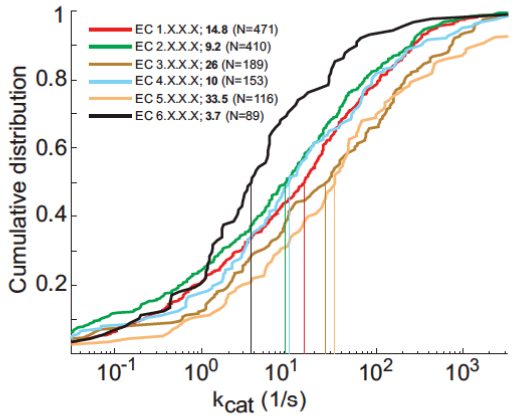


Figure S5

Enzymes of different classes have significantly different kcat values. Isomerase enzymes (EC 5.X.X.X) are significantly faster as compared to any other EC classes with p-value<0.005, except hydrolases (EC 3.X.X.X). Ligase enzymes (EC 6.X.X.X) are significantly slower as compared to any other EC classes with p-value<0.05. Bold numbers in the legend correspond to the median of each distribution, while numbers in parentheses represent the number of measured values in each distribution.

- Isomerases (EC 5.X .X .X) exhibit a median kcat of 33.5 s^{-1} , an order of magnitude higher than that of ligases (EC 6.X .X .X), with a median kcat of 3.7 s^{-1}
- Due to different reaction mechanisms and activation energy barriers

Effect of number of substrates in a reaction correlates with KM

- The higher the number of substrates, the lower the K_M for each substrate
  - as the number of substrates increases, lower K_M values are required to obtain the same concentration of the enzyme substrate complex

Figure S6

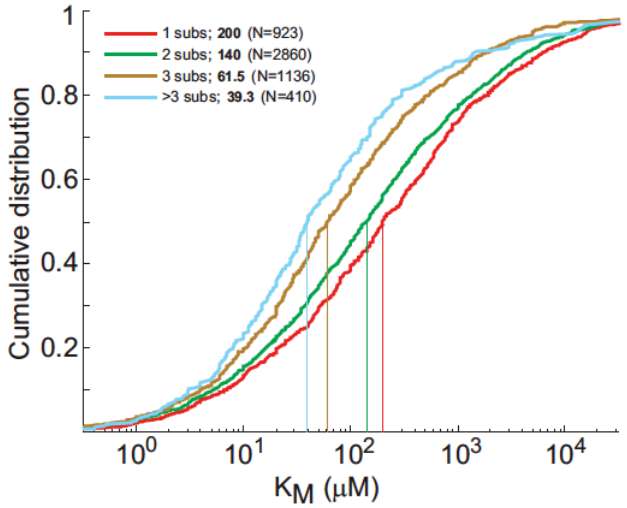


Figure S6

Reactions with a higher number of substrates are characterized by lower K_M values. The difference between any two groups is significant (p-value<0.01). Bold numbers in the legend correspond to the median of each distribution; in parentheses is the number of measured values in each distribution.

Are there significant correlation between various physicochemical properties of substrates and their KM values?

- Look at the following properties:

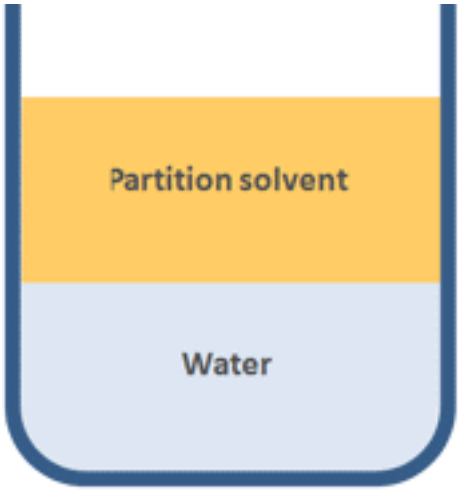
MW: molecular weight
LogP: log of octanol/water partition coefficient
SA: surface area
PSA: polar surface area;
NPS: non-polar surface area
CH: charge
ACH: absolute charge
NCA: number of charged atoms
HBD: hydrogen bond donors
HBT: hydrogen bond acceptors
HBT: total hydrogen bonds (acceptor + donors)
RB: rotatable bonds

- Keep in mind that associating K_M with substrate properties (and not enzyme-substrate interaction) may be misleading
  - ok with reactions with simple mechanism
  - Perhaps there is a correlation with the physiochemical properties

What is LogP?

- Lipophilicity represents the affinity of a molecule for a lipophilic (loves fats or lipids) environment.
- Measured by its distribution in a biphasic system (e.g. octanol, which is fatty and loves fat, and water)

$$\text{P = Partition Coefficient} = \frac{\text{Concentration dissolved in partition solvent}}{\text{Concentration dissolved in water}}$$



http://www.sirius-analytical.com/science/log-p/log-p-definitions

Plot MW vs K_M, and logP vs K_M

- To pick up a trend more clearly:
- The cyan area in Figure 3 was drawn as follows. For each data point, we examined the 150 point interval centered around that point (75 points on each side). We calculated the mean X-value and K_M for the window as well as the standard error (standard deviation) around the mean K_M.
- This process was repeated for each data point.
- Plotted are the standard deviation as a function of the mean X-value calculated this way.

For Small Substrates, K_M Decreases with Increasing Substrate Molecular Mass and Hydrophobicity

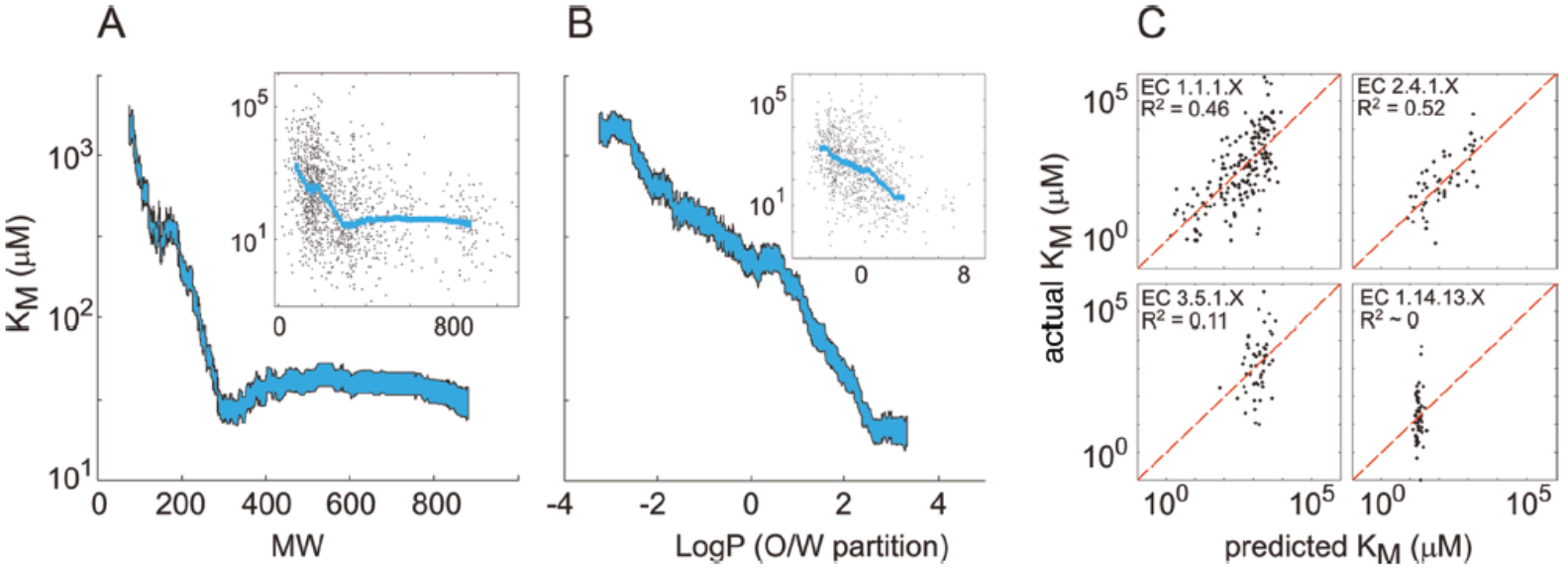


Figure 3. Correlations of K_M and substrate molecular mass and LogP. (A and B) The cyan area corresponds to the standard error around the mean value of a window centered at each X value (Supporting Information). The insets display the distribution of all the points. (A) Two regimes are noted: molecular mass of <350 Da, where the correlation between K_M and molecular mass was found to be negative ($R^2 = 0.13$; $p < 10^{-6}$; slope = -6×10^{-3} Da^{-1}) (Supporting Information), and molecular mass of >350 Da, where K_M values plateau at ~40 μM. (B) K_M vs LogP. Only small substrates (<350 Da) were used for this analysis ($R^2 = 0.24$; $p < 10^{-6}$; slope = 0.25 × $LogP^{-1}$) (Supporting Information). (C) Correlations between the predicted K_M, according to a substrate molecular mass and LogP, and the reported K_M values for several EC classes: 1.1.1.X ($R^2 = 0.46$), oxidoreductases, acting on the CH—OH group of donors, with NAD^+ or $NADP^+$ as the acceptor; 2.4.1.X ($R^2 = 0.52$), hexosyltransferases; 3.5.1.X ($R^2 = 0.11$), hydrolases, acting on carbon—nitrogen bonds, other than peptide bonds, in linear amides; and 1.14.13.X ($R^2 \sim 0$), oxygenases, acting on two donors, where NADH or NADPH is one donor, and incorporating one atom of oxygen into the other donor.

- K_M decreases, on average, by almost 2 orders of magnitude with an increasing molecular mass up to~350 Da
- K_M values also decrease ~100-fold with an increasing LogP for small substrates
- Are K_M and LogP are correlated?

Are KM and LogP are correlated?

Table S3 Correlations between substrate physiochemical properties and the kinetic parameters, for substrate with MW≤350

R^2		Molecular weight	Log of octanol/ water partition coefficient	Surface area	Polar surface area	Non-polar surface area	Charge	Absolute charge	Number of charged atoms	Hydrogen bond donors	Hydrogen bond acceptors	Total hydrogen bonds (acceptor + donors)	Number of rotatable bonds
k_{cat} (N=496)		0.03	0.02	0	0.02	0.01	0.02	0.02	0	0	0.01	0	0
K_M (N=943)		0.13	0.24	0.07	0.02	0.15	0	0	0.02	0.01	0.01	0.01	0
k_{cat}/K_M (N=490)		0.02	0.08	0.03	0	0.04	0.02	0.02	0	0.01	0	0	0
N=949	MW	1	0.01	0.36	0.16	0.19	0.01	0.02	0.01	0.15	0.28	0.29	0.2
	LogP	0.01	1	0.04	0.29	0.31	0	0	0.12	0.22	0.22	0.29	0
	SA	0.36	0.04	1	0.22	0.71	0	0.01	0.01	0.04	0.05	0.06	0.11
	PSA	0.16	0.29	0.22	1	0.01	0.14	0.17	0.29	0.27	0.62	0.6	0.13
	NPS	0.19	0.31	0.71	0.01	1	0.03	0.02	0.05	0.01	0.06	0.04	0.02
	CH	0.01	0	0	0.14	0.03	1	0.56	0.27	0.05	0.25	0.06	0.03
	ACH	0.02	0	0.01	0.17	0.02	0.56	1	0.51	0.01	0.17	0.06	0.08
	NCA	0.01	0.12	0.01	0.29	0.05	0.27	0.51	1	0	0.12	0.04	0.2
	HBD	0.15	0.22	0.04	0.27	0.01	0.05	0.01	0	1	0.28	0.66	0
	HBA	0.28	0.22	0.05	0.62	0.06	0.25	0.17	0.12	0.28	1	0.85	0.08
	HBT	0.29	0.29	0.06	0.6	0.04	0.06	0.06	0.04	0.66	0.85	1	0.05
	RB	0.2	0	0.11	0.13	0.02	0.03	0.08	0.2	0	0.08	0.05	1

Symbols:

- Perform correlation analysis
- No correlation between MW and LogP
- Non-polar surface area vs KM seems significant
  - stems from positive correlation between NPSA and both MW and LogP

Implications of the study

- Catalytic efficiencies well below diffusion rate
  - Can modify sequences to increase catalytic efficiency in synthetic bio applications
- Causes for low catalytic efficiencies:
  - A function of how they evolved and where the enzymes operate (e.g. primary vs secondary metabolism)
  - and/or a function of physiochemical constraints

Some open questions

- Kinetics within EC classes display significantly different behavior.
  - Kinetics within an RClass or within Reaction Modules?
    - Are they consistent with their metabolic pathway classification?
    - What's the interplay with physiochemical properties of substrates within class/ modules
- Are any of the kinetic parameters correlated with the reaction's transition states?
- What determines rate of productive collisions between enzymes and substrates?
  - Median kcat/KM = ~10^5 M^{-1} s^{-1}
  - Diffusion rate = 10^8 M^{-1} s^{-1}
  - On interplay average, 1 in 1000 molecules go through a productive collision.

What about enzymes acting on non-natural substrates?

- Prevailing assumption:
  - Enzymes are assumed "specific": they operate on "one" product (the natural product)
  - EC classification originally assigned one reaction per enzyme

- Changing view:
  - Enzymes can catalyze more than one reaction due to "promiscuity"
  - Over one third of Escherichia coli 's enzymes catalyze two or more transformations {Nam, Lewis, Lerman, .. 2012}.
  - Now EC numbers list additional reactions catalyzed by the same reaction

Promiscuity boosted adaptive evolution

- Jensen's hypothesis:
  - Ancient cells have possessed small gene content
  - Primitive enzymes may have possessed broad specificity and undeveloped regulation mechanisms, offering maximum biochemical flexibility with minimal gene content.
  - Gene duplication, with accumulated mutations, provided the opportunity for increased gene content and increased specialization of the diverging enzymes
  - Introduction of a single new enzyme might have created new multi-step pathways
  - Substrate specialization was further reinforced by the development of regulatory mechanisms.

Jensen, R. A. "**Enzyme Recruitment in Evolution of New Function.**" *Annual review of microbiology* 30 (1976): 409–25.

Enzyme Promiscuity – Two main types

- Two types:
  - Substrate promiscuity, where an enzyme can bind to multiple substrates and exhibits broad specificity, and
  - Catalytic promiscuity, where an enzyme catalyzes a different reaction with a different transition state
- In some cases, an enzyme exhibits both behaviors (e.g., tyrosine phosphatase isoform δ protein)
- Focus has been on substrate promiscuity

Example Promiscuous activity in *E. coli*

Native reaction in *E. coli*

Malonyl CoA + NADPH + H^+ $\xleftrightarrow{\text{Malonyl-CoA Reductase (MCR)}}$ Malonic semialdehyde + $NADP^+$ + CoA

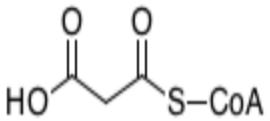
KEGG reaction

C00083 + C00005 + C00080 $\xleftrightarrow{\text{R00740}}$ **C00222** + C00006 + **C00010**

Putative reaction due to MCR promiscuity

Succinyl CoA + NADPH + H^+ ($\mathbf{C00091}$) $\xleftrightarrow{\text{Malonyl-CoA Reductase (MCR)}}$ Succinic semialdehyde + $NADP^+$ + CoA ($\mathbf{C00232}$)

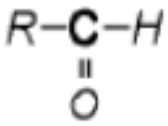
COMPOUND: C00083

Entry	C00083 Compound
Name	Malonyl-CoA; Malonyl coenzyme A
Formula	C24H38N7O19P3S
Exact mass	853.1156
Mol weight	853.5803
Structure	 C00083 Mol file KCF file DB search Jmol KegDraw
Reaction	R00233 R00353 R00740 R00742 R00743 R00744 R00989 R01613 R01614 R01626 R02224 R02447 R02482 R02505 R04050 R04356 R04386 R04618 R04709 R05188 R05190 R05331 R06448 R06453 R06458 R06459 R06480 R06481 R06482 R06483 R06510 R06568 R06625 R06635 R06637 R06641 R06643 R06644 R06645 R06796 R06797 R06799 R06800 R06801 R06817 R06821 R07250 R07251 R07253 R07255 R07719 R07721 R07731 R07741 R07754 R07757 R07758 R07884 R07885 R07886 R07910 R07938 R07987 R07988 R07989 R08796 R08797 R08805 R08806 R09090 R09195 R09258 R09419 R09527 R09621 R10171 R10173 R10233 R10234 R10238 R10239 R10485 R10825 R10960 R10965 R10967 R11124 R11125 R11412 R11435 R11445 R11516 R11584 R11587 R11588 R11615 R11616 R11664 R11667 R11671 « hide

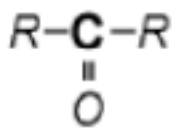
RCLASS: RC00184

Help

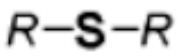
Entry	RC00184 RClass
Definition	C4a-C5a:*-S2a:C1b+O4a-C1b+O5a
	
Reactant pair	C00083_C00222 C00091_C00232 C00100_C00479 C00136_C01412 C00154_C00517 C00593_C19685 C00609_C02843 C00609_C03371 C00609_C20683 C04076_C05535 Path search



C4a (350)

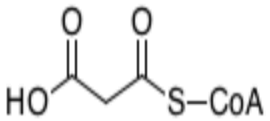


C5a (3595)



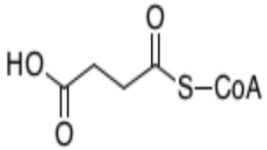
S2a (420)

COMPOUND: C00083

Entry	C00083 Comp
Name	Malonyl-CoA; Malonyl coenzyme A
Formula	C24H38N7O19P3S
Exact mass	853.1156
Mol weight	853.5803
Structure	 C00083 Mol file KCF file DB search Jmol
Reaction	R00233 R00353 R00740 R00742 R00 R01614 R01626 R02224 R02447 R02 R04386 R04618 R04709 R05188 R05 R06458 R06459 R06480 R06481 R06 R06625 R06635 R06637 R06641 R06643 R06644 R06645 R06796 R06797 R06799 R06800 R06801 R06817 R06821 R07250 R07251 R07253 R07255 R07719 R07721 R07731 R07741 R07754 R07757 R07758 R07884 R07885 R07886 R07910 R07938 R07987 R07988 R07989 R08796 R08797 R08805 R08806 R09090 R09195 R09258 R09419 R09527 R09621 R10171 R10173 R10233 R10234 R10238 R10239 R10485 R10825 R10960 R10965 R10967 R11124 R11125 R11412 R11435 R11445 R11516 R11584 R11587 R11588 R11615 R11616 R11664 R11667 R11671 « hide

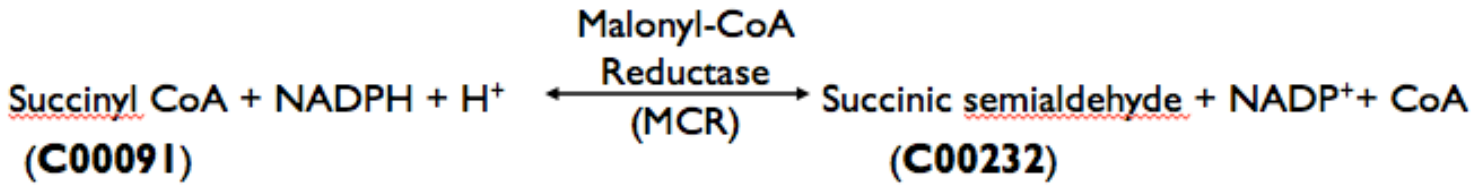
COMPOUND: C00091

Help

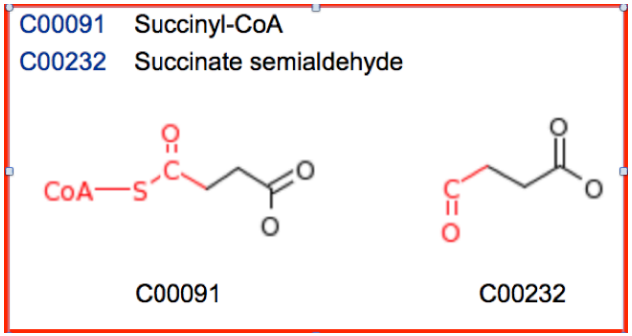
Entry	C00091 Compound
Name	Succinyl-CoA; Succinyl coenzyme A
Formula	C25H40N7O19P3S
Exact mass	867.1313
Mol weight	867.6069
Structure	 C00091 Mol file KCF file DB search Jmol KegDraw
Reaction	R00265 R00405 R00406 R00407 R00410 R00432 R00727 R00829 R00830 R00832 R00833 R01197 R01559 R01777 R01780 R02084 R02407 R02570 R02990 R03154 R04365 R05587 R05588 R06904 R06908 R08549 R09280 R10343 R10640 R10641 R10904 R11773
Pathway	map00020 Citrate cycle (TCA cycle)

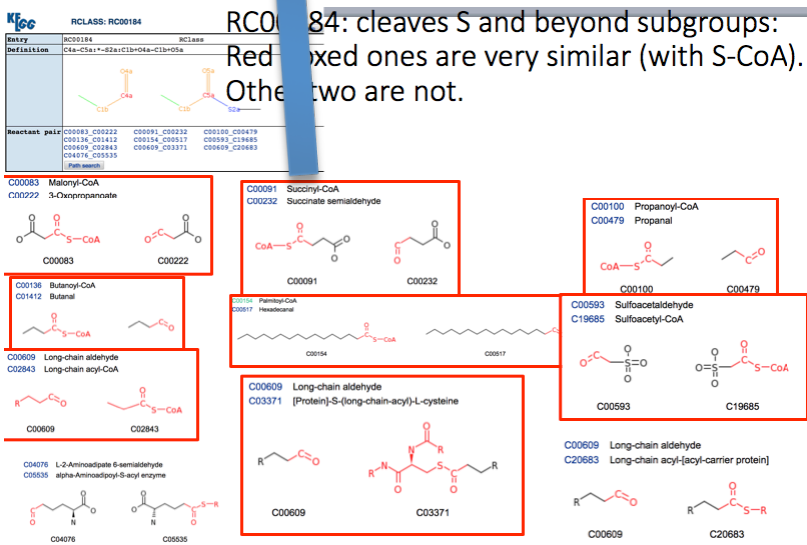
Is there a known reaction that performs the promiscuous operation?

Putative reaction due to MCR promiscuity



- Reactant Pair





RC0 84: cleaves S and beyond subgroups: Red xed ones are very similar (with S-CoA). Othe two are not.

- Associated with Reaction R09280
- Is it in *E. coli*?
  - No!!
- Is it catalyzed by MCR?
  - Nope. By Enzyme 1.2.1.76 (succinate-semialdehyde dehydrogenase)
- Question: are MCR and 1.2.176 functionally similar? Don't know!!

Predicting substrate promiscuity – some techniques

1. Substrate similarity between a query molecule and the native substrate that is known to be catalyzed by the enzyme

- The GEMP-Path algorithm determines if there exists in BRENDA a reaction that displays the desired promiscuous activity in terms of cofactor and substrate similarity [31].
- SimZyme suggests possible relationships between an enzyme and a metabolite based on its similarity of other metabolites that the enzyme acts upon [37].
- Use binding site covalence and thermodynamic favorability to score promising enzyme candidates (in context of building synthesis pathways) [30].

Predicting substrate promiscuity – some techniques

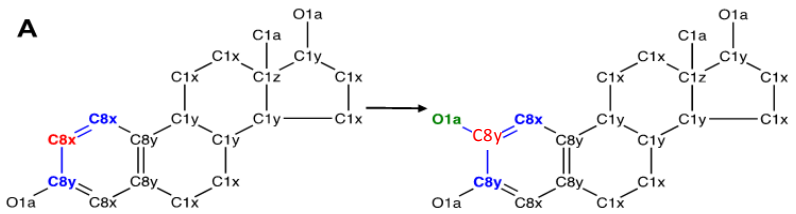
2. Machine Learning techniques (Support Vector Machines, SVMs) either to:
   - predict the promiscuity of a given enzyme [38] or
   - predict if a particular enzyme will transform a particular substrate of interest [32]

3. Analyzing kinetic parameter of enzymes
   - J-index, does not seem generalizable based on today's paper[39]

Predicting outcomes of promiscuous transformations

- If the enzyme operates on a non-natural substrate, what is the product (or derivative)?
- Techniques:
  - Rule-based, example BNICE: hand curated set of rules based on examining ECs up to their 3rd level of specificity [41]
  - PROXIMAL – operators derived from RPAIRS with neighbor's neighbor data [22]

A



B **Lookup Table**

RPAIR	Key (reactant RDM pattern)			Value (product RDM pattern)	
	R	M1	M2	R	D
RP00390	C8x	C8x	C8y	C8y	O1a
RP03118	C4a	C1c	O4a	C6a	O6a
RP11918	C8x	N4y	N5x	C8y	O5x
RP11919	N4x	C8x	C8y	N4y	C1a

References

22. Yousofshahi, M., Manteiga, S., Wu, C., Lee, K., and Hassoun, S. "**PROXIMAL: A Method for Prediction of Xenobiotic Metabolism**" *(submitted) Journal of Chemical Information and Modeling* (2014)

30. Cho, A., Yun, H., Park, J. H., Lee, S. Y., and Park, S. "**Prediction of Novel Synthetic Pathways for the Production of Desired Chemicals.**" *BMC systems biology* 4 (2010): 35.

31. Campodonico, M. A., Andrews, B. A., Asenjo, J. A., Palsson, B. O., and Feist, A. M. "**Generation of an Atlas for Commodity Chemical Production in Escherichia Coli and a Novel Pathway Prediction Algorithm, GEM-Path.**" *Metabolic engineering* 25 (2014): 140–58.

32. Pertusi, D. A., Moura, M. E., Jeffryes, J. G., Prabhu, S., Walters Biggs, B., and Tyo, K. E. J. "**Predicting Novel Substrates for Enzymes with Minimal Experimental Effort with Active Learning**" *Metabolic Engineering* 44, no. August (2017): 171–181.

37. Pertusi, D. A., Stine, A. E., Broadbelt, L. J., and Tyo, K. E. J. "**Efficient Searching and Annotation of Metabolic Networks Using Chemical Similarity.**" *Bioinformatics (Oxford, England)* 31, no. 7 (2015): 1016–24.

38. Carbonell, P. and Faulon, J.-L. "**Molecular Signatures-Based Prediction of Enzyme Promiscuity.**" *Bioinformatics (Oxford, England)* 26, no. 16 (2010): 2012–9.

39. Nath, A. and Atkins, W. M. "**A Quantitative Index of Substrate Promiscuity**" *Biochemistry* 47, no. 1 (2008): 157–166.

41. Finley, S. D., Broadbelt, L. J., and Hatzimanikatis, V. "**Computational Framework for Predictive Biodegradation.**" *Biotechnology and bioengineering* 104, no. 6 (2009): 1086–97.

Open Questions

- There is no global study on kinetic parameters for non-natural parameters
- How do you generate a balanced "reaction" that depicts the promiscuous operation?
  - Useful in creating synthetic pathways or analyzing metabolic disruption due to enzyme promiscuity

Homework #2

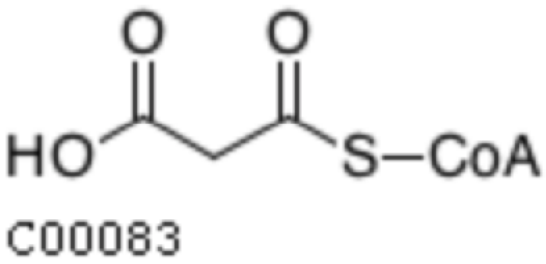
- Enzymes are assumed promiscuous
- PROXIMAL is used to identify some candidate substrates
- Data

DataSet Num	Reaction	Reference Molecule	Candidate Substrates
1	R02946	C00810	C00022,C00026,C00111,C00118,C00149,C00188,C00197,C00199,C00231,C00257,C00258,C00318,C00332,C00345,C04411
2	R01976	C00332	C00022,C00026,C00091,C00111,C00199,C00231,C05223
3	R01175	C00877	C00083,C00091,C00100,C00332
4	R01172	C00136	C00083,C00100,C00332

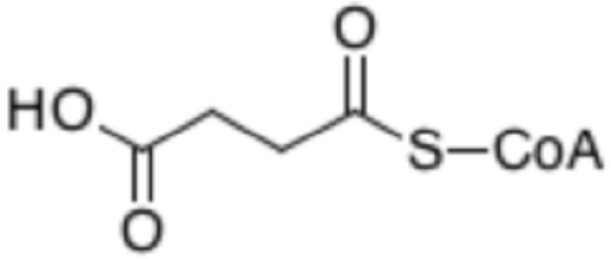
Homework #2 – part 1

Explore enzyme promiscuity via molecular similarity

- Let's use fingerprints to determine enzyme promiscuity towards a new set of substrates (call them candidates).
- Let's assess the similarity between natural substrate and candidates
- For example: how similar are the natural substrate (Malonyl CoA C00083) for MCR vs the candidate substrate (Succinyl CoA , C00091)



C00083



C00091

Let's use rdkit in python

- History
  - 2000-2006: Developed and used at Rational Discovery for building predictive models for Tox, biological activity
  - June 2006: Open-source (BSD license) release of software, Rational Discovery shuts down
  - to present: Open-source development continues, use within Novartis, contributions from Novartis back to open-source version
- Functional overview:
  - Input/output: standard formats: SMILES, SMARTS, MOL, FASTA, …
  - Fingerprinting: Daylight-like, atom pairs, topological torsions, Morgan algorithm "MACCS keys", extended reduced graphs, etc.
  - 2D pharmacophores (per web site, not the fastest, best, …)
  - Many many more….
- # Recommend using python 2.7

To generate Fingerprints and similarity scores

1. Generate mol files for each of your molecules (from smiles, for example)

```
ms = [Chem.MolFromSmiles('CCOC'), Chem.MolFromSmiles('CCO')]
```

2. Call fingerprints on each mol file

```
fps = [FingerprintMols.FingerprintMol(x) for x in ms]
```

3. Compute fingerprint similarity using metric of your choice:

```
a= DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.TanimotoSimilarity)
e=  DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.DiceSimilarity)
```

4. Print the scores:

```
print ('Tanimoto vs Dice using defaults', a, e)
```

The code

```
# need to import these two things
from rdkit import DataStructs
from rdkit.Chem.Fingerprints import FingerprintMols

ms = [Chem.MolFromSmiles('CCOC'), Chem.MolFromSmiles('CCO')]
fps = [FingerprintMols.FingerprintMol(x) for x in ms]
a= DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.TanimotoSimilarity)
e=  DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.DiceSimilarity)
print ('Tanimoto vs Dice using default keys', a, e)
```

To generate other fingerprints

```
from rdkit.Chem import MACCSkeys
fps = [MACCSkeys.GenMACCSKeys(x) for x in ms]
a= DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.TanimotoSimilarity)
e=  DataStructs.FingerprintSimilarity(fps[0],fps[1],
metric=DataStructs.DiceSimilarity)
print ('Tanimoto vs Dice using MACCSkeys
fingerprints', a, e)
```

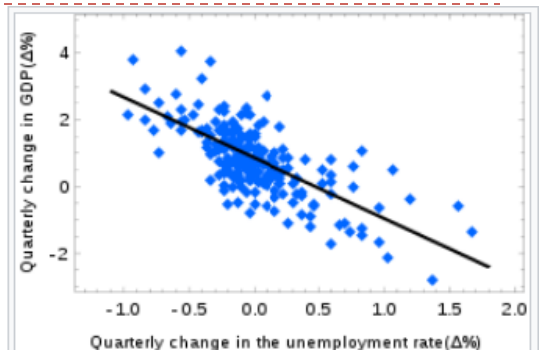
Homework #2 – part 2

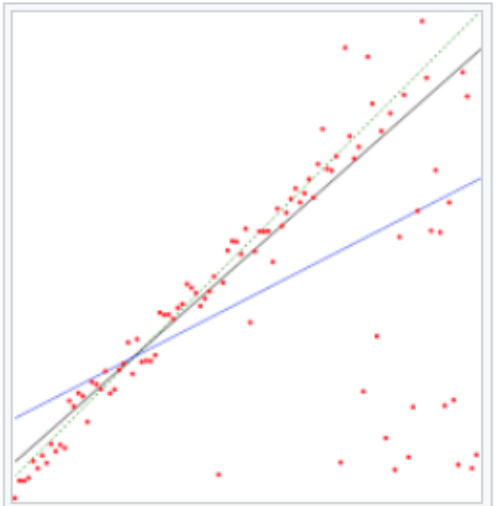
- Use BRENDA to look up functional parameters for the enzymes that catalyze the reactions in the data sets.
- Look at the data for both natural enzymes and for candidate substrates, and compare against average values we studied today

https://en.wikipedia.org/wiki/Coefficient_of_determination

What are R^2 values?

- https://www.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/r-squared-or-coefficient-of-determination
- Coefficient of determination, denoted R^2 or r^2 and pronounced "R squared", is the proportion of the variance in the dependent variable (Y) that is predictable from the independent variable (X).
- Used with a "statistical model" that either predicts future outcomes or tests hypotheses





Comparison of the Theil–Sen estimator (black) and simple linear regression (blue) for a set of points with outliers. Because of the many outliers, neither of the regression lines fits the data well, as measured by the fact that neither gives a very high R^2.

What is the p-value

- https://www.khanacademy.org/math/statistics-probability/significance-tests-one-sample/tests-about-population-mean/v/hypothesis-testing-and-p-values
- There is always a Null Hypothesis: usually assumes the observed data has the same distribution as the model
  - The summary metric (e.g. mean) is the same between the two
- There is an Alternate Hypothesis: opposite of the Null.
- p-value: if the null-hypothesis is true, what's the likelihood of getting the observed data
  - If p-value is small, this is an extreme condition, and unlikely happening by chance. Reject Null Hypothesis
  - If p-value is large, then ya, observed data in line with the model. Accept Null Hypothesis
- p-value calculation shown is only applicable when you have normal distributions, and not very good with big data sets – so sometimes need to come up with alternate way of computing p-value

Calculating the p-value in the "Moderately Efficient Enzyme" paper

- Calculate *p-values* for observed R^2 using a Monte Carlo permutation test (also known as approximate permutation test or random permutation test)
- The *p-value* corresponds to the null assumption that $R^2 = 0$ and was calculated by shuffling the Y-values, randomly assigning them to X-values, and calculating the resulting R^2.
- This process was repeated 10^6 times.
- Using the resulting distribution of R^2 values, the p-value was calculated as the fraction of times where the randomized R^2 values were equal or higher than the original R^2.