

Structural and functional constraints on protein evolution

Claus O. Wilke

The University of Texas at Austin

... P N G E N R K - N I E K F T E K K N F F Y R L K K F I E K N F - S P N D F V I L L M G D I N V A I N D K D I G I S E N N ...
... P Q G E N R N - N K I K F N K K K N F Y K N L I K Y V Q K K I - F E N K N I I I L M G D M N I S P E D Q D I G I D P Y H ...
... P H G E S F Y - K T D K F E E K K F F Y Q K L Y L F L K R N C - E K N A H I L I M G D M N I S P T D L D I G I L S V N S ...
... P N G E S K K - N L K K F E I K K I F Y K S L F L F L N K F Y - E K N M H I L I M G D M N I S P E D L D V G I L S L T S ...
... P Q G K S I D -- H P D Y Q A K H R F F D R L L N L F Q R E F - S P H T P L L W V G D M N V A P T D I D V ----- T S ...
... P N G N P A P -- G P K F D Y K L R W F E R L R L R A Q E L I - A T G A P V V I A G D Y N V M P T E L D V ----- Y K ...
... P N G N P A P -- G P K F D Y K L R W F D R L I T H A Q G L L - A A G K P V I L L T G D F N V M P T E L D V ----- Y K ...
... P N G N P V E -- T P K Y P Y K L R W M D R L I R Y A E D R L - A L E E P L V L A G D Y N V L P T P D D V ----- A N ...
... P N G N P P Q -- T E K Y P Y K L K W M D R L L A Y S K E R L - K S E E P F V L A G D F N V I P T P E D V ----- Y N ...
... P N G N P P N -- T E K Y P Y K L K W M S R L R D Y A R E R L - K T E E P L I L A G D F N V I P A A A D V ----- S N ...
... P N G N P V G -- S E K Y P Y K L S W M A R L R D Y A Q Q R L - K T E E P L I L A G D F N V I P Q A E D V ----- H N ...
... P N G N P V P -- G P K Y D Y K L A W M E R L R A R A I E L L - K S E A P F V M A G D Y N I I P Q P M D A ----- A K ...
... P N G N P A P -- G P K Y D Y K L A W M A R M H A R V E S L L - P L E E P L V F C G D Y N V I P Q A E D A ----- A K ...
... P N G N P A P -- G P K Y D Y K L A W M E R L E A R A R E E L L - A E E M P A L M A G D Y N V I P Q A E D A ----- A R ...
... P N G N P A P -- G P K Y D Y K L A W M E R L R A R A E A L L - K A E E P A L M A G D Y N V I P Q A E D A ----- A K ...
... P N G N P A P -- G P K F D Y K L A W M Q R L E A R A K A L L - A D E M P F I L M A G D Y N I I P Q A E D A ----- A K ...
... P N G N P V D -- T E K F S Y K L E W M D R L I A R A K E L L - L L E E P F V M M G D Y N I I P H E D D V ----- H D ...
... P N G N P I D -- S D K F P Y K L S W M E R L R S R V K E L L - T Y E E P F V V A G D Y N V I P T P E D V ----- Y D ...
... P N G N P L G -- T D K F P Y K L A W M D R L R R H A A L R L - A E E Q P F L L L G D Y N V I P E P K D A ----- R N ...
... P N G N P L G -- T E K F P Y K L R W M D R L I A H A R T R L - A E E T P F L L L G D Y N V I P E P K D A ----- R N ...
... P N G N P L G -- T E K F P Y K L G W M D R L I A H A K R R L - D D E I P Y I L L G D Y N V I P D P M D A ----- K N ...
... P N G N P V S A D S V K F P Y K L G W M E R L E A W A Q E R L - E L E E P L I L A G D Y N V I P M P V D C ----- H D ...
... P N G N P V D -- T E K F P Y K L R W M E R L Q A F A E D R L - A L E E P L V L A G D Y N V I P E P V D C ----- H N ...

completely conserved

PNGENRK-NIEKFTEKKNFFYRLKKFIEKNF-SPNDVFVLLM**GDINVAINDKD**I**GISENN**...
POGENRN-NKIKFNKKKNFYKNLIK**VOKKT**-FENKNTTT**MGDMNT**SPEDODIGIDPYH...
PHGESFY-KTDKFEEKKFFYQKLYLF**Y**
PNGESKK-NLKKFEI**K**KKI**F**YKSLFLF**L**
PQGKSID--HPDYQA**K**HRF**D**RLLNL**F**QREF-SPHTPLLWVG**GD**M**N**VAPTDIDV---TS...
PNGNPAP--GPKFDY**K**LRWF**E**RLRLRA**Q**ELI-ATGAPVVIAG**D**YN**V**MPTE**I**LDV---YK...
PNGNPAP--GPKFDY**K**LRWF**D**RLL**I**THA**Q**GLL-AAGKPVLLTG**DF**N**V**MPTE**I**LDV---YK...
PNGNPVE--TPKYPY**K**LRWM**D**RLL**I**RYA**E**DR**L**-ALEEPLVLAG**D**YNVLPTPDDV---AN...
PNGNPPQ--TEKYPY**K**LKWMDRLL**A**YSKERL-KSEEPFVLAG**DF**N**V**IP**T**PEDV---YN...
PNGNPPN--TEKYPY**K**LKWMSRLRD**Y**ARERL-KTEEPLILAG**DF**NVIPAAADV---SN...
PNGNPVG--SEKYPY**K**L**S**WMARLRD**Y**AQQRL-KTEEPLILAG**DF**NVIPQAEDV---HN...
PNGNPVP--GPKYDY**K**LAWMERLRARAI**E**LL-KSEAPFVMA**G**DYN**I**IP**Q**PM**D**A---AK...
PNGNPAP--GPKYDY**K**LAWMAR**M**HARVES**S**LL-PL**E**PLVFC**G**DYNVIPQAEDA---AK...
PNGNPAP--GPKYDY**K**LAWMERLE**A**RARE**E**LL-AEEMPALMA**G**DYNVIPQAEDA---AR...
PNGNPAP--GPKYDY**K**LAWMERLRARAE**A**LL-KAEEPALMA**G**DYNVIPQAEDA---AK...
PNGNPAP--GPKFDY**K**LAWM**Q**RLEARAK**A**LL-AD**E**MPFL**M**A**G**DYN**I**IP**Q**AEDA---AK...
PNGNPVD--TEKFSY**K**LEWMDRLL**I**ARAK**E**LL-L**LE**EPFVMM**G**DYN**I**IP**H**EDDV---HD...
PNGNPID--SDKFPY**K**L**S**W**M**ER**L**RSRV**K**ELL-TYEEPFVVA**G**DYNVIP**T**PEDV---YD...
PNGNPLG--TDKFPY**K**LAWMDRLL**R**RA**A**RL**L**-A**E**EQPFL**LLL**
PNGNPLG--TEKFPY**K**LRWF**D**RLL**I**AHART**R**LL-A**E**ETPFL**LLL**
PNGNPLG--TEKFPY**K**LGWMDRLL**I**AHAK**R**RL-D**D**EIPY**LLL**
PNGNPVSADS**V**KFPY**K**LGW**M**ER**L**EA**W**A**Q**ER**L**-E**L**EEPL**L**ILA**G**DYNVIP**M**P**V**DC---HD...
PNGNPVD--TEKFPY**K**LRW**M**ER**L****Q**A**F**A**E**DR**L**-A**L**EEPLVLA**G**DYNVIP**E**P**V**DC---HN...

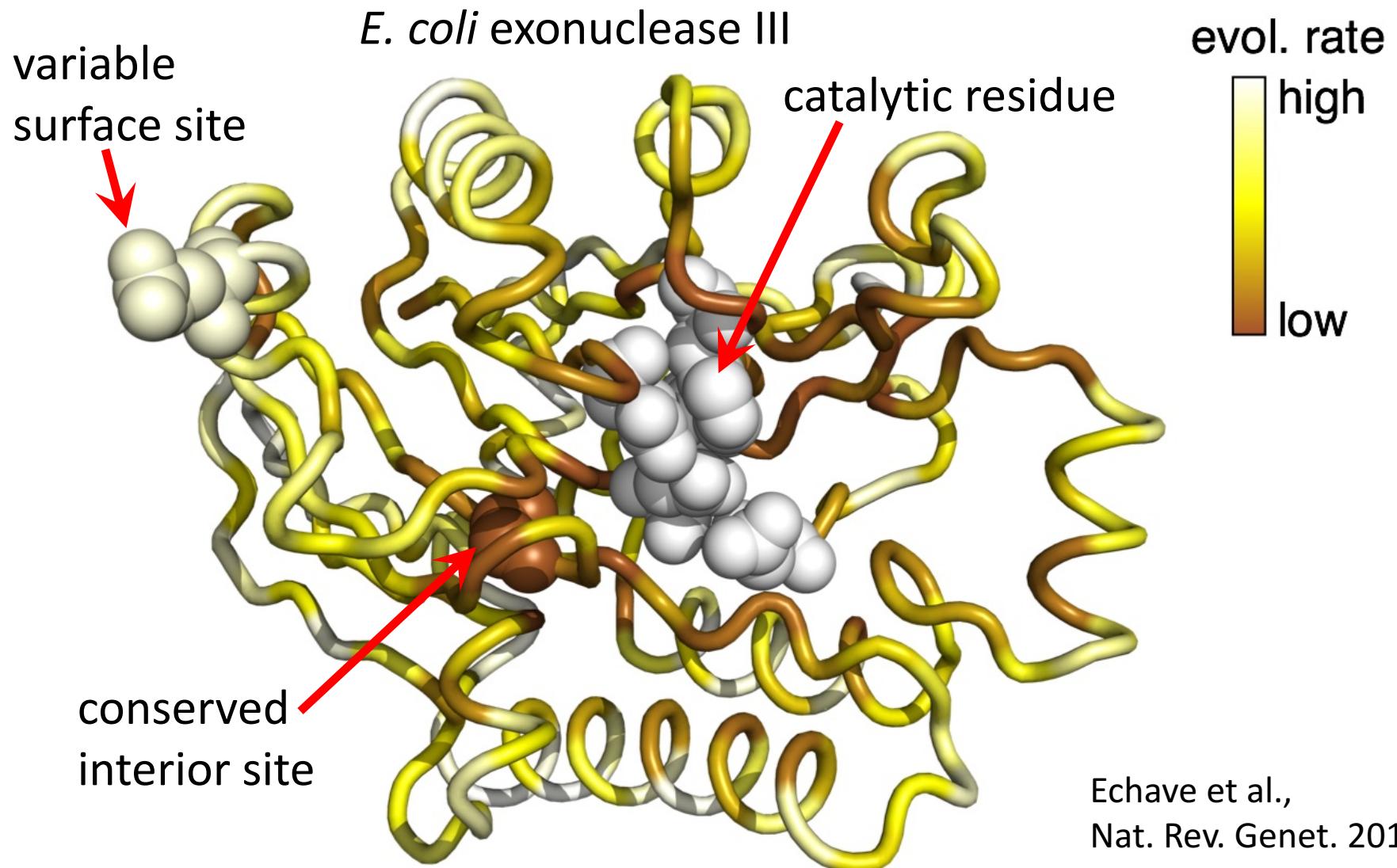
mostly conserved

... P N G E N R K - N I E K F T E K K N F F Y R L K K F I E K N F - S P N D F V I L M G D I N V A I N D K D I G I S E N N ...
... P Q G E N R N - N K I K F N K K K N F Y K N L I K Y V O K K T - F E N K N T T I M G D M N T S P E D Q D I G I D P Y H ...
... P H G E S F Y - K T D K F E E K K F F Y Q K L Y L F ...
... P N G E S K K - N L K K F E I K K I F Y K S L F L F L N K F Y - E K N M H I L I M G D M N I S P E D L D V G I S L T S ...
... P Q G K S I D -- H P D Y Q A K H R F F D R L L N L F Q R E F - S P H T P L L W V G D M N V A P T D I D V - - - - T S ...
... P N G N P A P -- G P K F D Y K L R W F E R L R L R A Q E L I - A T G A P V V I A G D Y N V M P T E L D V - - - - Y K ...
... P N G N P A P -- G P K F D Y K L R W F D R L I T H A Q G L L - A A G K P V L L T G D F N V M P T E L D V - - - - Y K ...
... P N G N P V E -- T P K Y P Y K L R W M D R L I R Y A E D R L - A L E E P L V L A G D Y N V L P T P D D V - - - - A N ...
... P N G N P P Q -- T E K Y P Y K L K W M D R L L A Y S K E R L - K S E E P F V L A G D F N V I P T P E D V - - - - Y N ...
... P N G N P P N -- T E K Y P Y K L K W M S R L R D Y A R E R L - K T E E P L I L A G D F N V I P A A A D V - - - - S N ...
... P N G N P V G -- S E K Y P Y K L S W M A R L R D Y A Q Q R L - K T E E P L I L A G D F N V I P Q A E D V - - - - H N ...
... P N G N P V P -- G P K Y D Y K L A W M E R L R A R A I E L L - K S E A P F V M A G D Y N I I P Q P M D A - - - - A K ...
... P N G N P A P -- G P K Y D Y K L A W M A R M H A R V E S L L - P L E E P L V F C G D Y N V I P Q A E D A - - - - A K ...
... P N G N P A P -- G P K Y D Y K L A W M E R L E A R A R E E L L - A E E M P A L M A G D Y N V I P Q A E D A - - - - A R ...
... P N G N P A P -- G P K Y D Y K L A W M E R L R A R A E A L L - K A E E P A L M A G D Y N V I P Q A E D A - - - - A K ...
... P N G N P A P -- G P K F D Y K L A W M Q R L E A R A K A L L - A D E M P F L M A G D Y N I I P Q A E D A - - - - A K ...
... P N G N P V D -- T E K F S Y K L E W M D R L I A R A K E L L - L L E E P F V M M G D Y N I I P H E D D V - - - - H D ...
... P N G N P I D -- S D K F P Y K L S W M E R L R S R V K E L L - T Y E E P F V V A G D Y N V I P T P E D V - - - - Y D ...
... P N G N P L G -- T D K F P Y K L A W M D R L R R H A A L R L - A E E Q P F L L L G D Y N V I P E P K D A - - - - R N ...
... P N G N P L G -- T E K F P Y K L R W M D R L I A H A R T R L - A E E T P F L L L G D Y N V I P E P K D A - - - - R N ...
... P N G N P L G -- T E K F P Y K L G W M D R L I A H A K R R L - D D E I P Y L L L G D Y N V I P D P M D A - - - - K N ...
... P N G N P V S A D S V K F P Y K L G W M E R L E A W A Q E R L - E L E E P L I L A G D Y N V I P M P V D C - - - - H D ...
... P N G N P V D -- T E K F P Y K L R W M E R L Q A F A E D R L - A L E E P L V L A G D Y N V I P E P V D C - - - - H N ...

quite variable

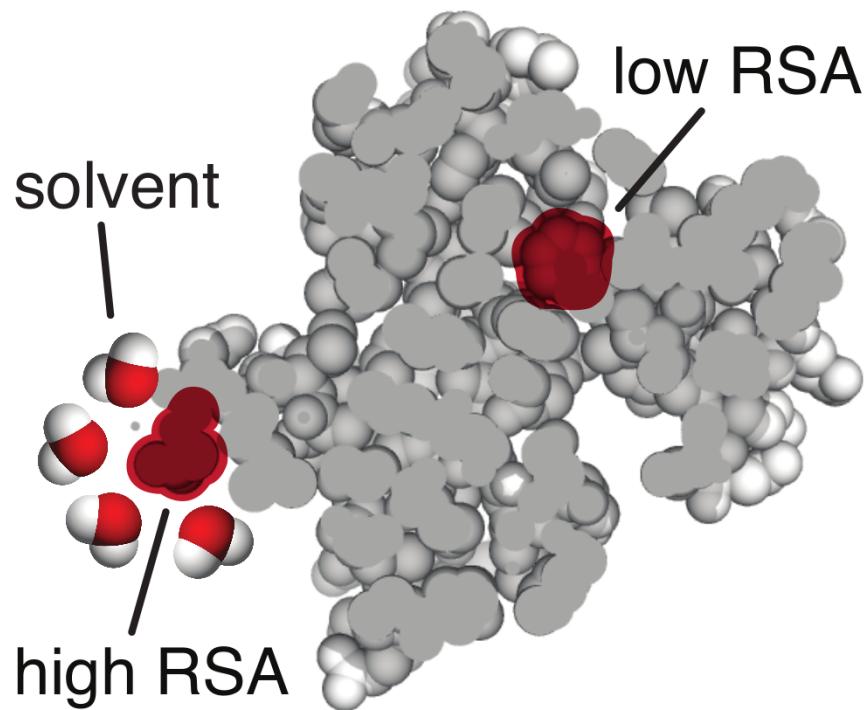
... PNGENRK - NIEKFTEKKNFFYRLKKFIEKNF - SPNDVFVLLMGDINVAINDKD**I**GISENN ...
... PQGENRN - NKIKFNKKKNFYKNL**I**IKYVOKKT FENKNTT**I**M**I**DMNISPEDQDIGIDPYH ...
... PHGESFY - KTDKFEEKKFFYQKL**I**YLF E**I**N**I**DMNISPTDLDIGLSVNS ...
... PNGESKK - NLKKFEI**I**KKIFYKSLFL**I**Y**I**E**I**N**I**DMNISPEDLDVG**I**SLTS ...
... PQGKSID - HPDYQAKHRFFDR**I**LLNL**I**F**I**QREF SPHTPLLWVG**I**DMNVAPTDIDV --- TS ...
... PNGNPAP - GPKFDYKLRWF**I**ERL**I**RLRA**I**QELI ATGAPVVIAGDYNVM**I**MPTELDV --- YK ...
... PNGNPAP - GPKFDYKLRWF**I**DR**I**LTHAQG**I**LL AAGKPVLLTGDFNV**I**MPTELDV --- YK ...
... PNGNPVE - TPKYPYKLRWM**I**DRL**I**RYAED**I**RL ALEEPLVLAGDYNVLPTPDDV --- AN ...
... PNGNPPQ - TEKYPYKLKWMDR**I**LLAYS**I**KERL KSEEPFVLAGDFN**I**PTPEDV --- YN ...
... PNGNPPN - TEKYPYKLKWMSR**I**LDYARERL KTEEPLILAGDFN**I**IPAAADV --- SN ...
... PNGNPVG - SEKYPYKLSW**I**MARL**I**RDYAQ**I**QL KTEEPLILAGDFN**I**IPQAEDV --- HN ...
... PNGNPVP - GPKYDYKLA**I**W**I**MERL**I**RARAI**I**LL KSEAPFVMAGDYN**I**IPQMDA --- AK ...
... PNGNPAP - GPKYDYKLA**I**W**I**M**I**ARM**I**HARV**I**ES**I**LL PLEEPLVFCGDYN**I**IPQAEDA --- AK ...
... PNGNPAP - GPKYDYKLA**I**W**I**MERL**I**EARARE**I**LL AEEMPALMAGDYN**I**IPQAEDA --- AR ...
... PNGNPAP - GPKYDYKLA**I**W**I**MERL**I**RARA**I**ALL KAEEPALMAGDYN**I**IPQAEDA --- AK ...
... PNGNPAP - GPKFDYKLA**I**W**I**MQ**I**RI**I**EARAK**I**ALL ADEM**I**FLMAGDYN**I**IPQAEDA --- AK ...
... PNGNPVD - TEKFPYKLEWMDR**I**LI**I**ARAK**I**ELL LLEEPFVMMGDYN**I**IPHEDDV --- HD ...
... PNGNPID - SDKFPYKLSW**I**MERL**I**RSRV**I**KELL TYEEPFVVAGDYN**I**IP**I**PEDV --- YD ...
... PNGNPLG - TD**I**KFPYKLA**I**W**I**MDR**I**LI**I**RR**I**HA**I**AL**I**RL AEEQPFL**I**LG**I**DYN**I**IP**I**PEPKDA --- RN ...
... PNGNPLG - TEKFPYKLRWM**I**DR**I**IA**I**HART**I**RL AEETPFL**I**LG**I**DYN**I**IP**I**PEPKDA --- RN ...
... PNGNPLG - TEKFPYKL**I**GWMDR**I**IA**I**HAK**I**RL DDEIPY**I**LL**I**GDYN**I**IP**I**DPMDA --- KN ...
... PNGNPVSAD SVKFPYKL**I**GW**I**MERL**I**EA**I**WA**I**Q**I**ER**I**RL ELEEPLILAGDYN**I**IP**I**MPVDC --- HD ...
... PNGNPVD - TEKFPYKLRW**I**MERL**I**Q**I**A**I**FA**I**ED**I**RL ALEEPLVLAGDYN**I**IP**I**EPVDC --- HN ...

Patterns of sequence variation make more sense in a structural context

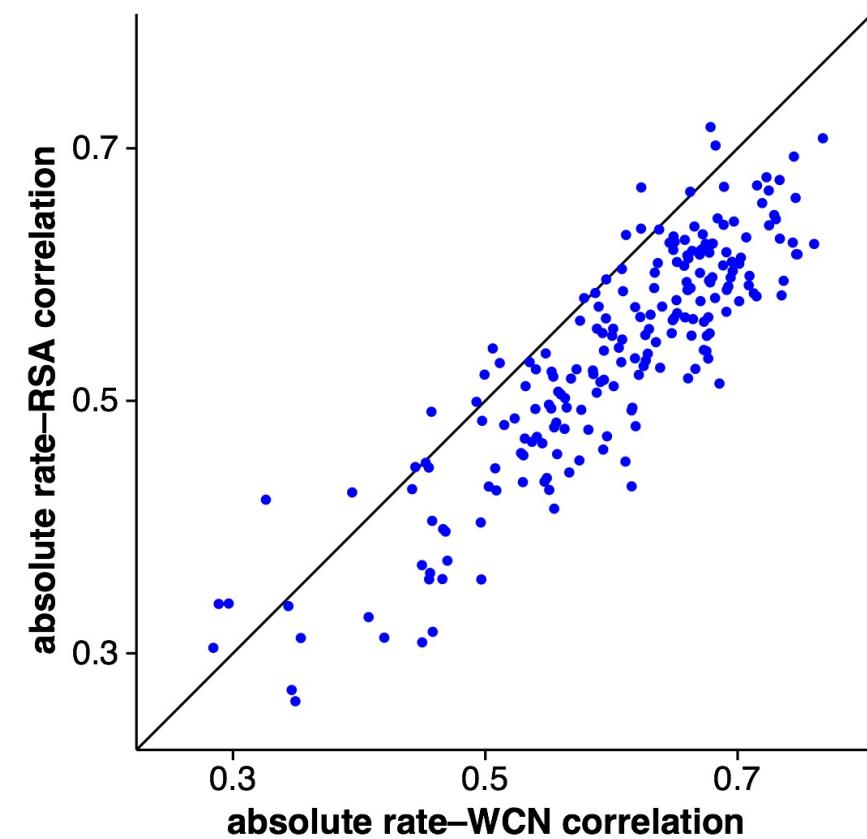


We measure structure with RSA and WCN

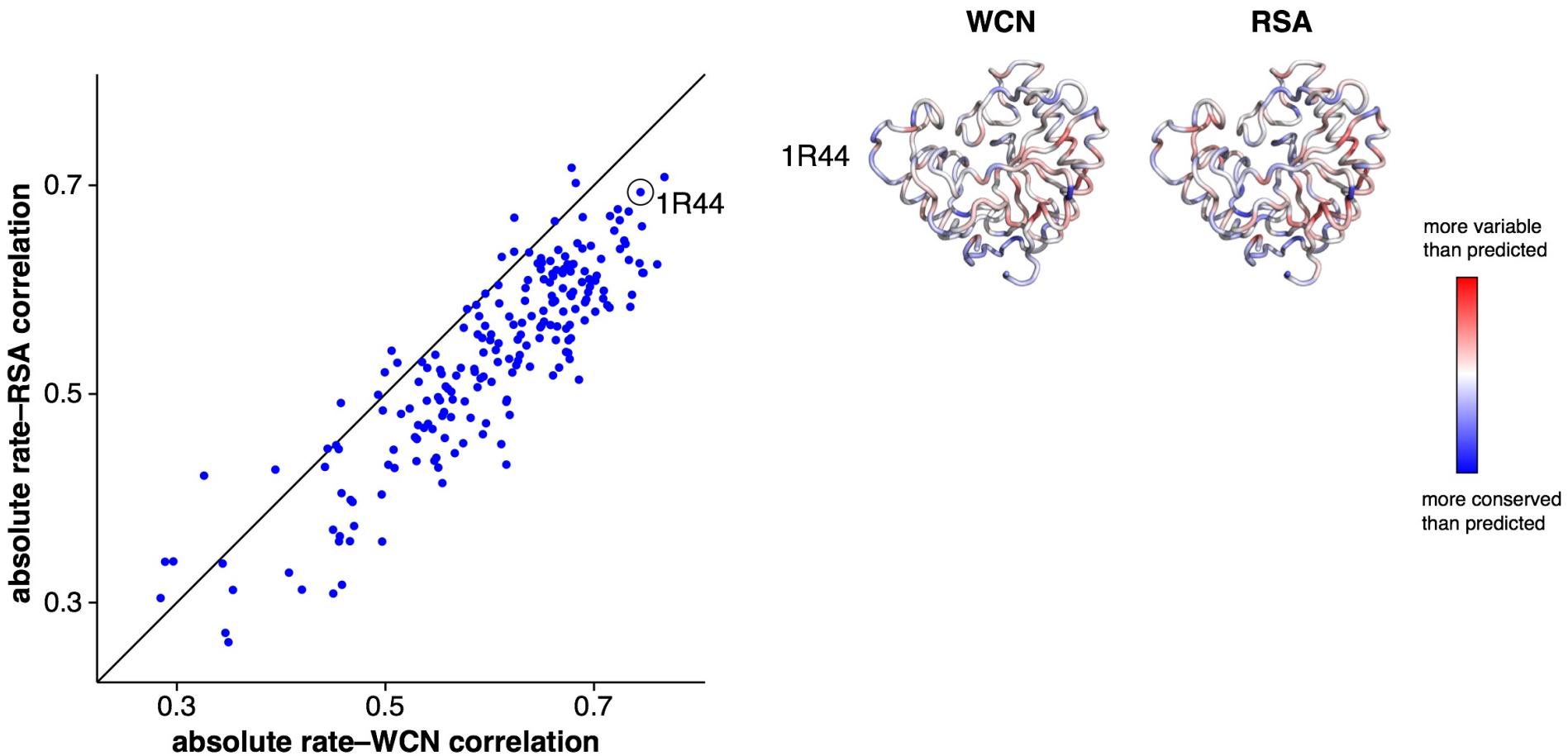
Relative Solvent Accessibility



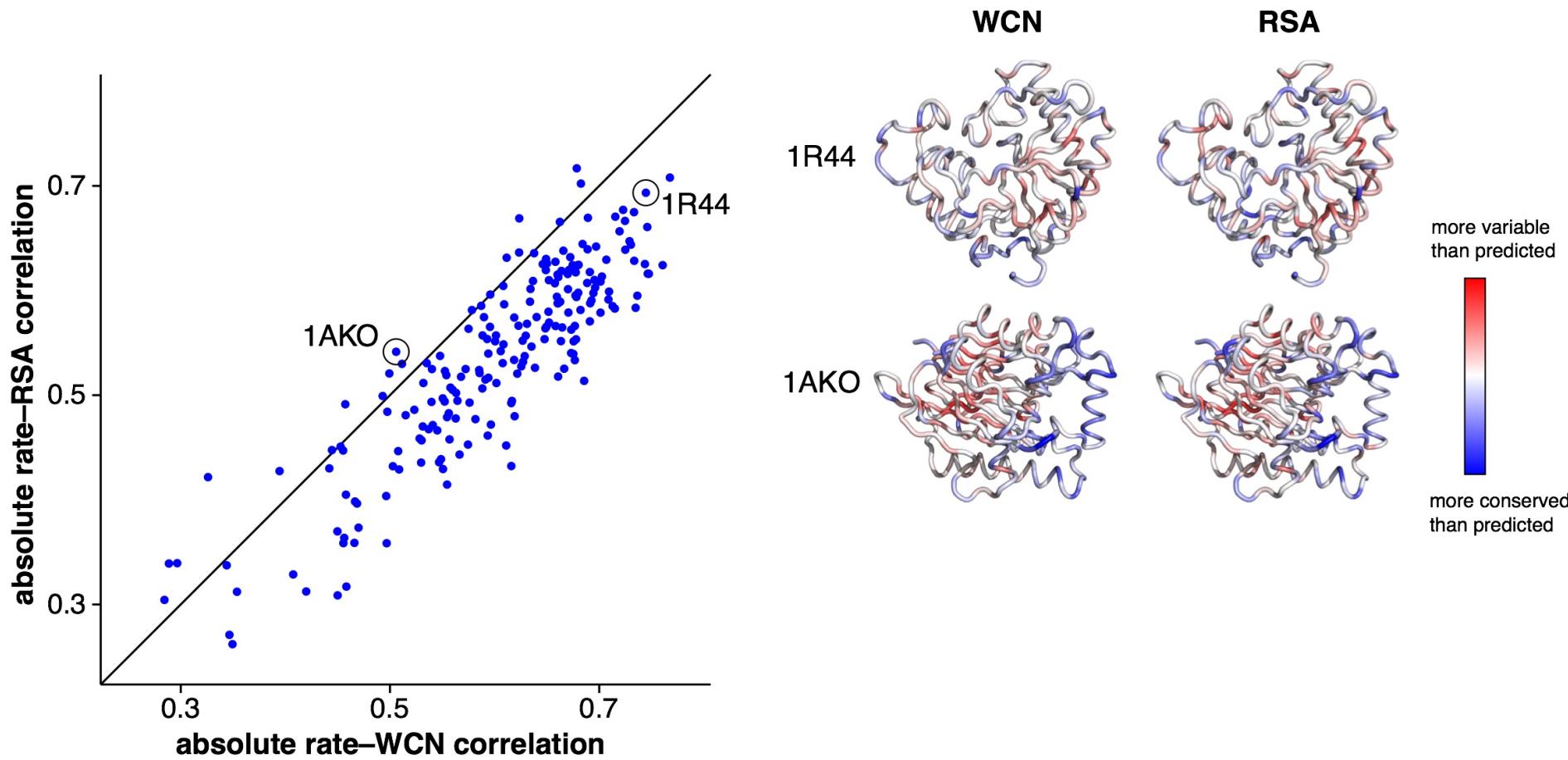
Structure is a good rate predictor for some proteins, but not so much for others



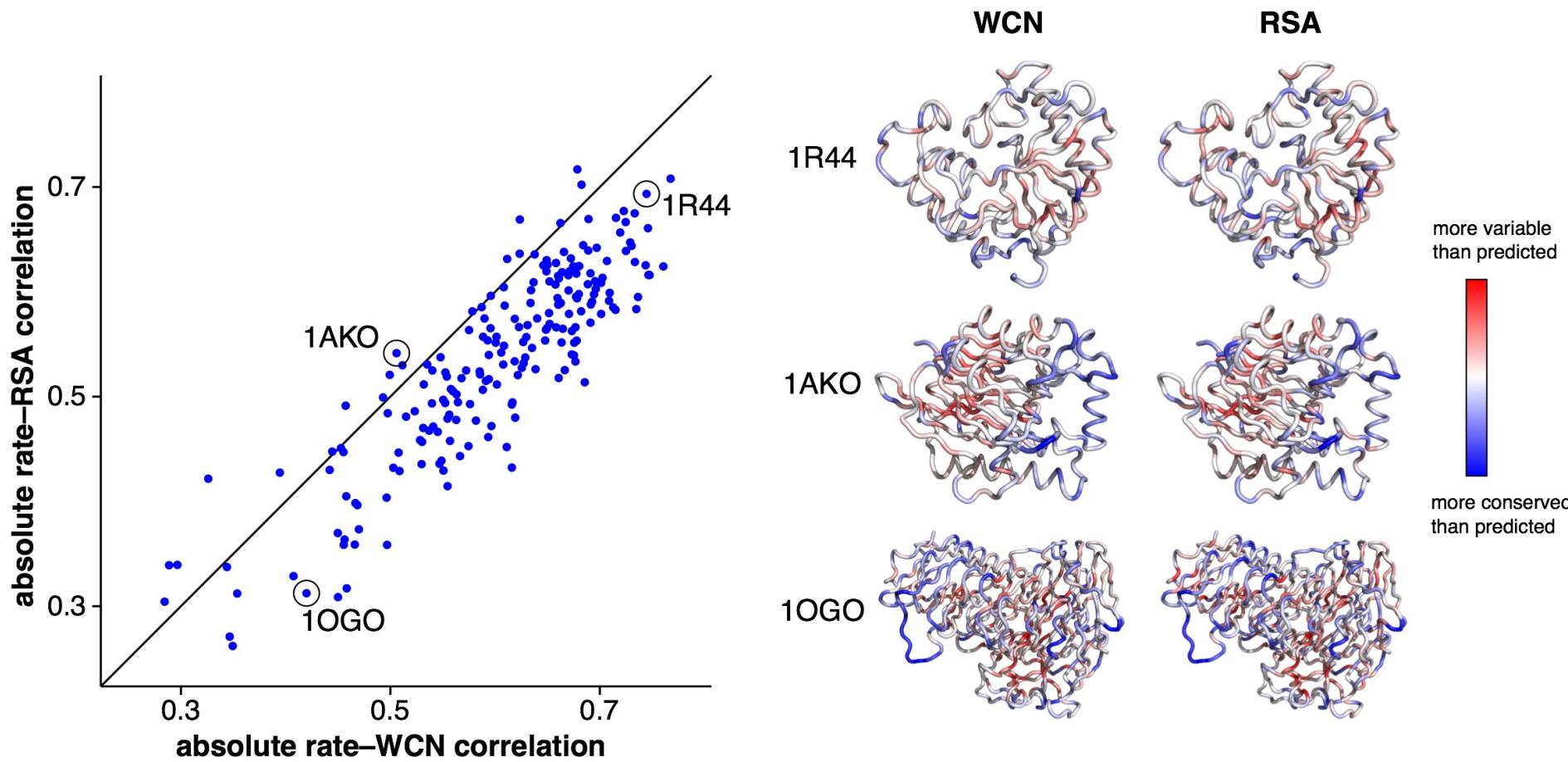
Structure is a good rate predictor for some proteins, but not so much for others



Structure is a good rate predictor for some proteins, but not so much for others



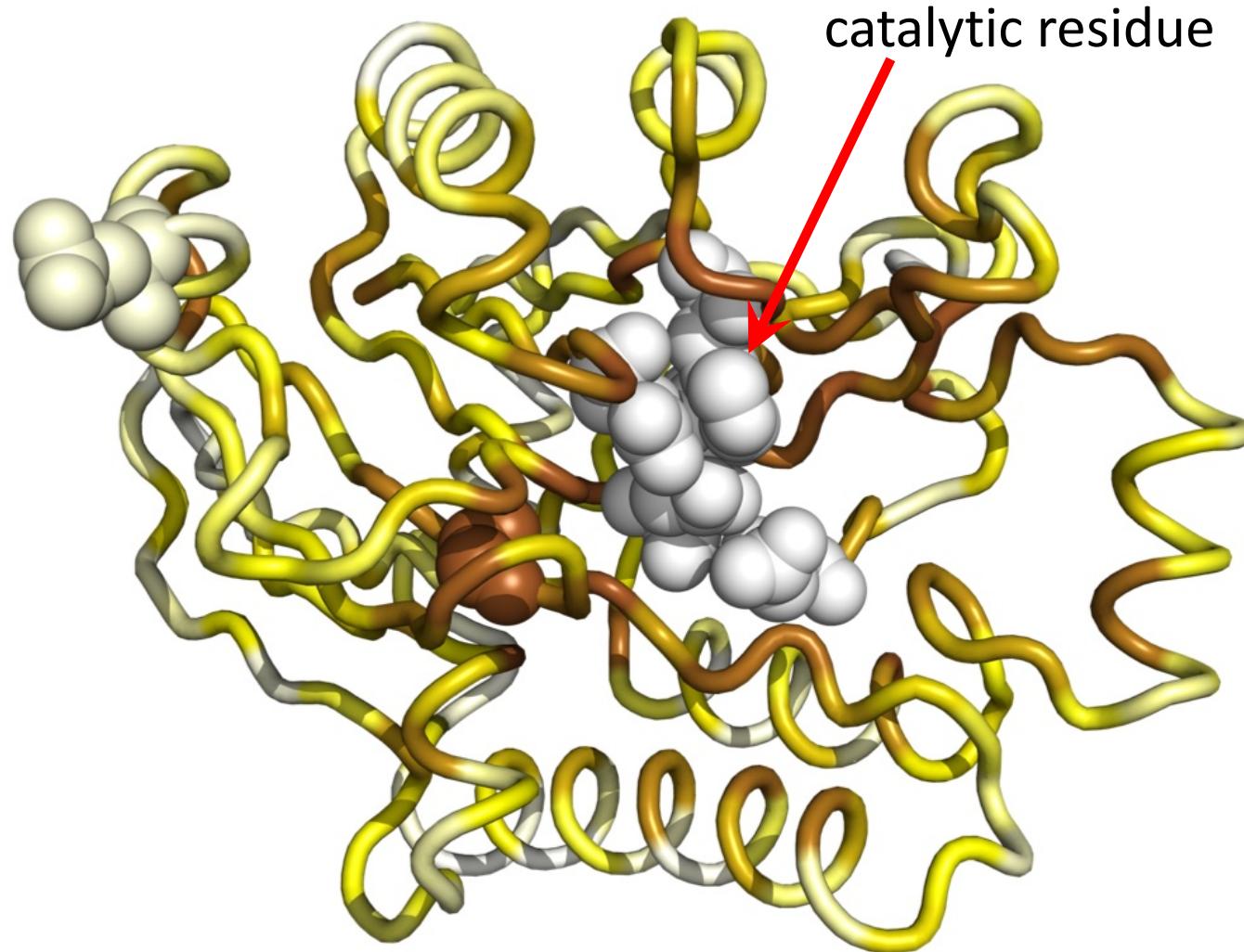
Structure is a good rate predictor for some proteins, but not so much for others



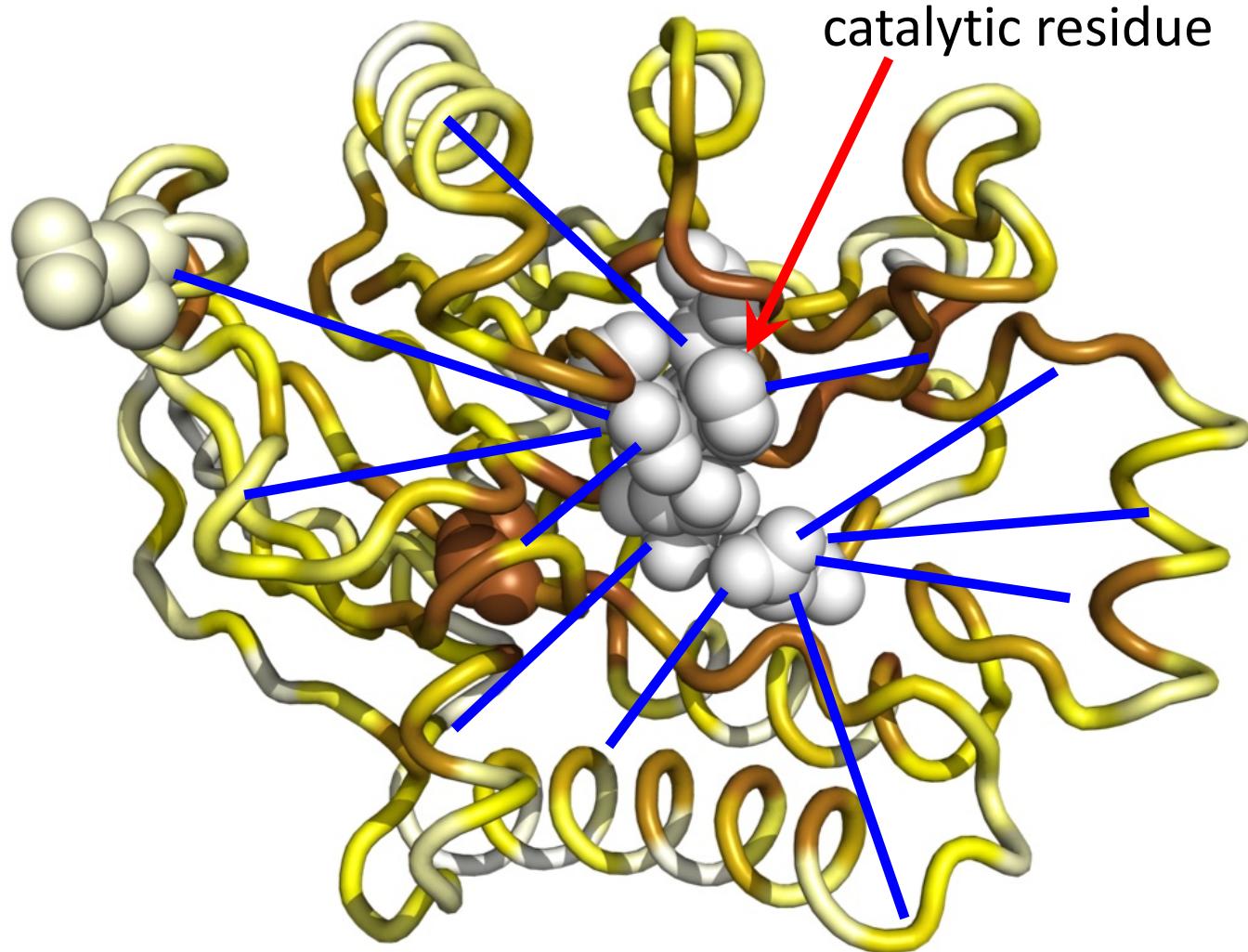
Part I: How does protein function constrain site-specific evolution?

Work by graduate student Ben Jack
In collaboration with Julian Echave

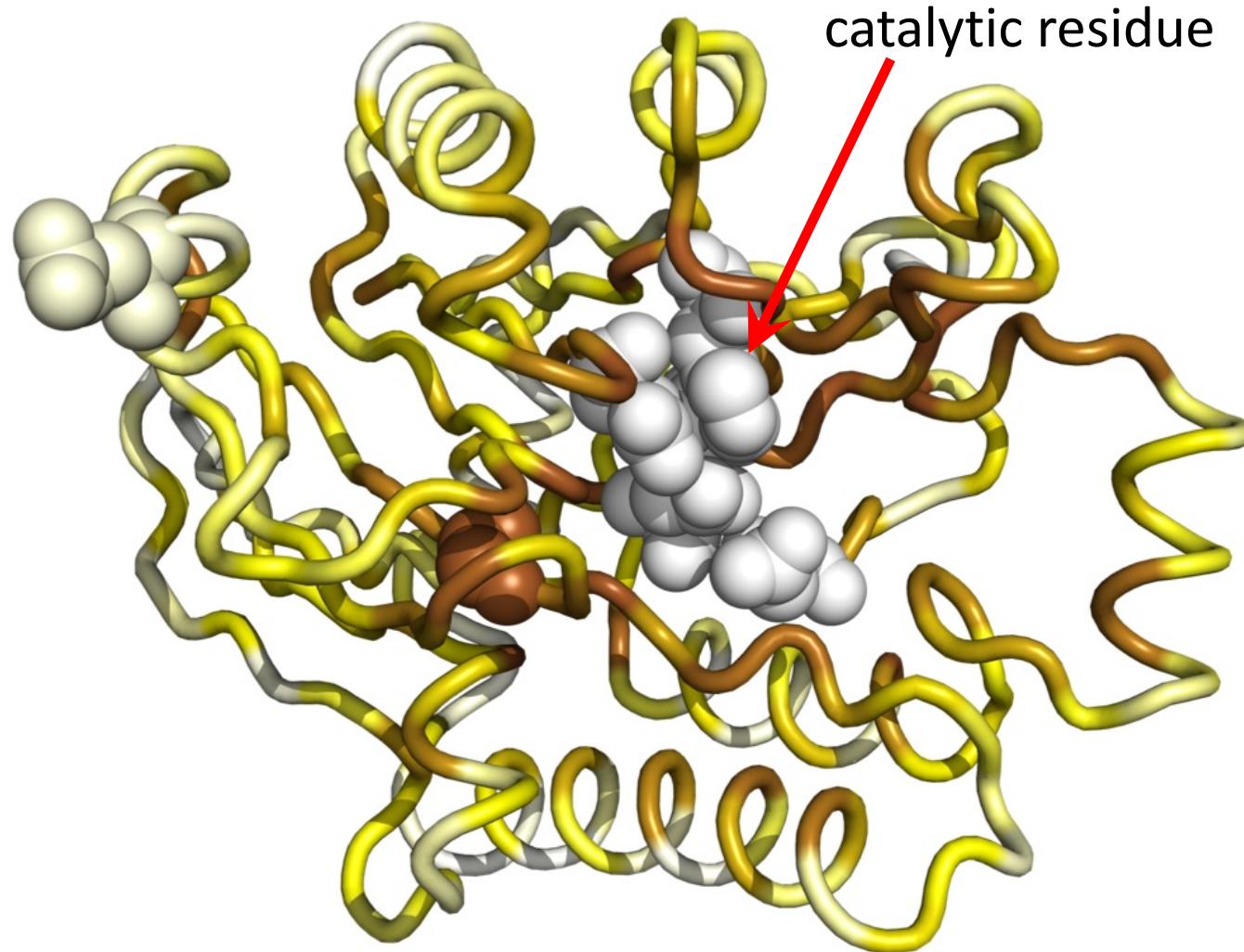
We measure function with distance to the nearest catalytic residue



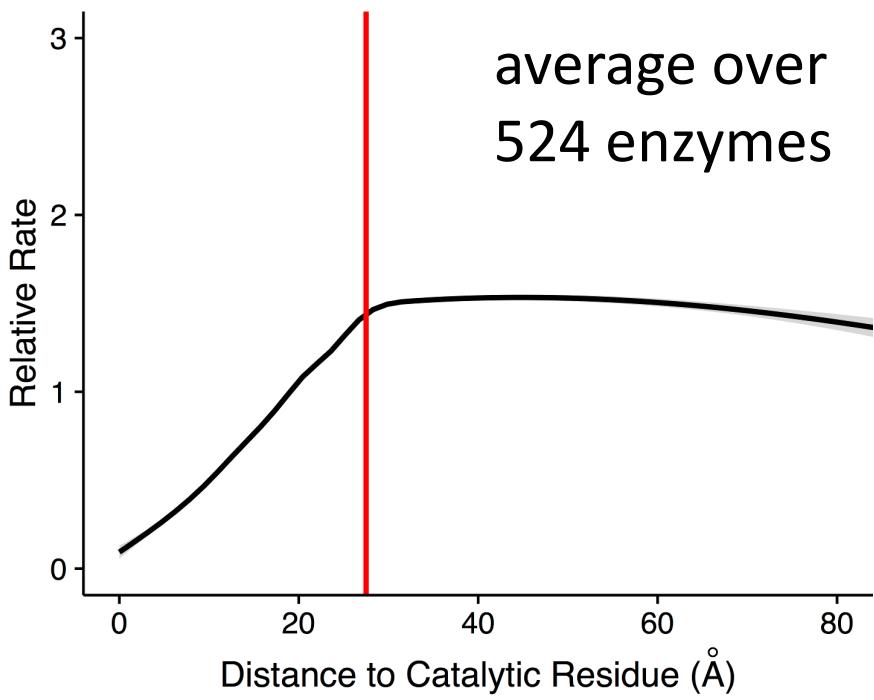
We measure function with distance to the nearest catalytic residue



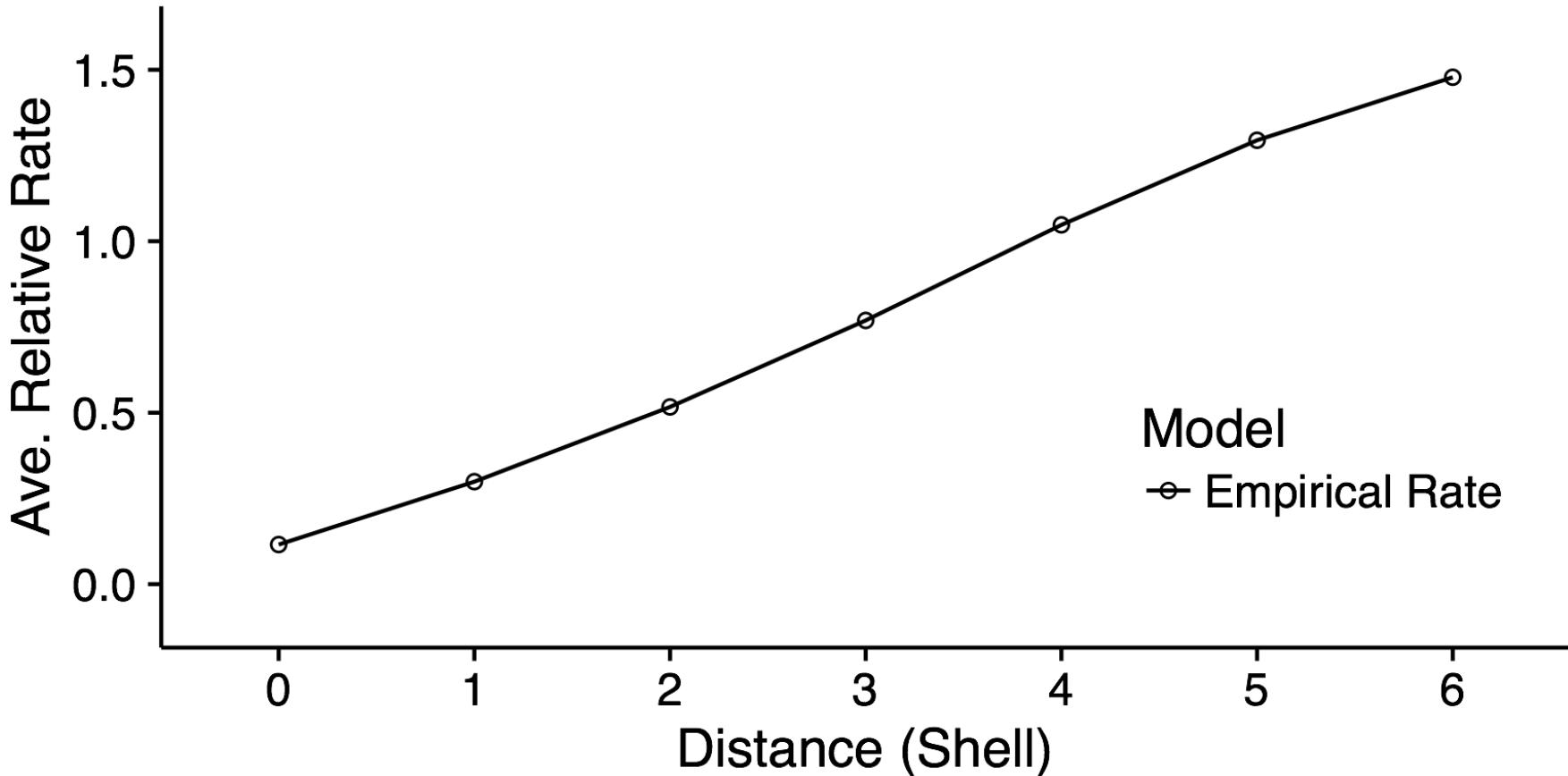
We measure function with distance to the nearest catalytic residue



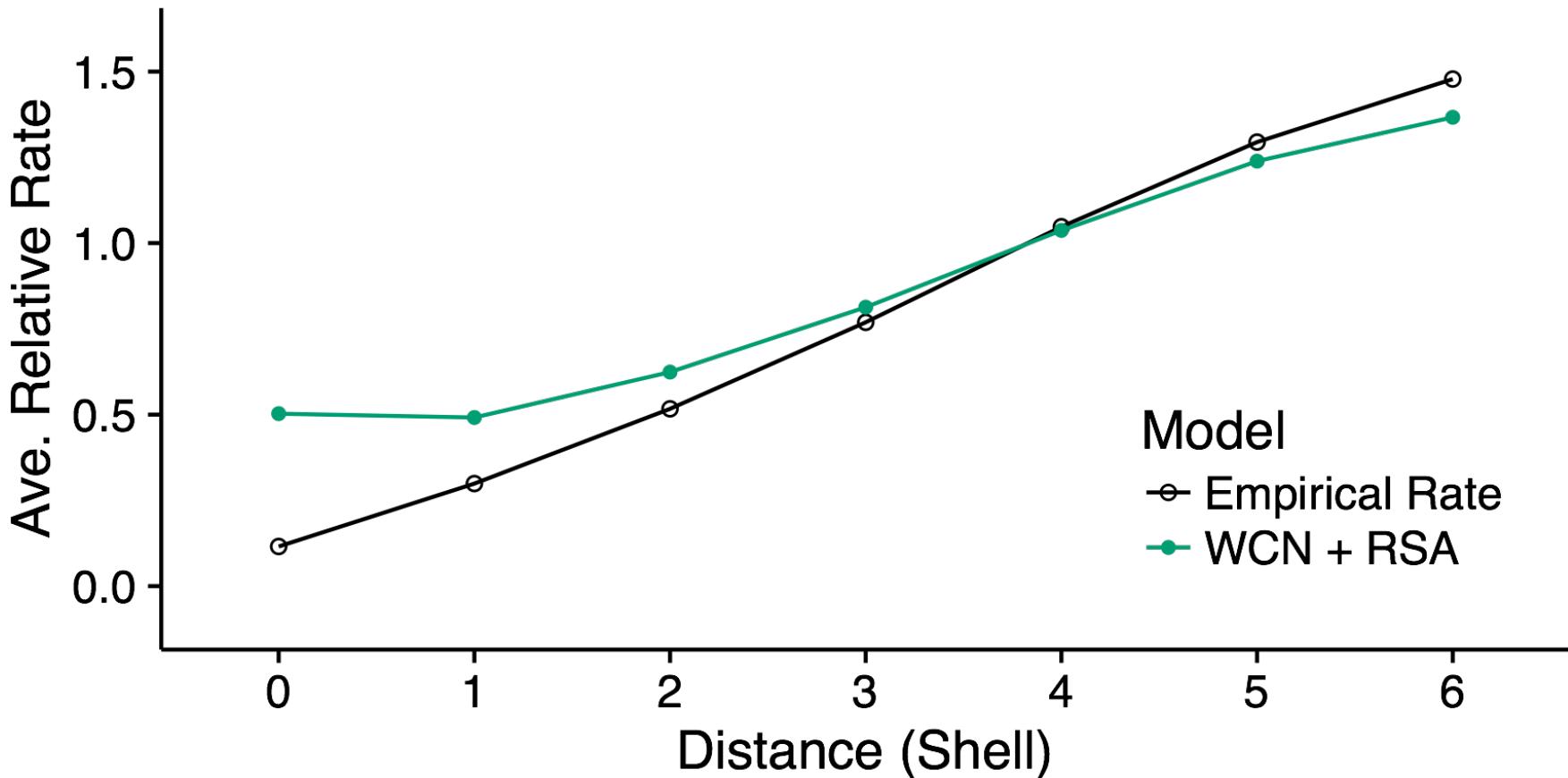
Catalytic residues impose long-range conservation gradients



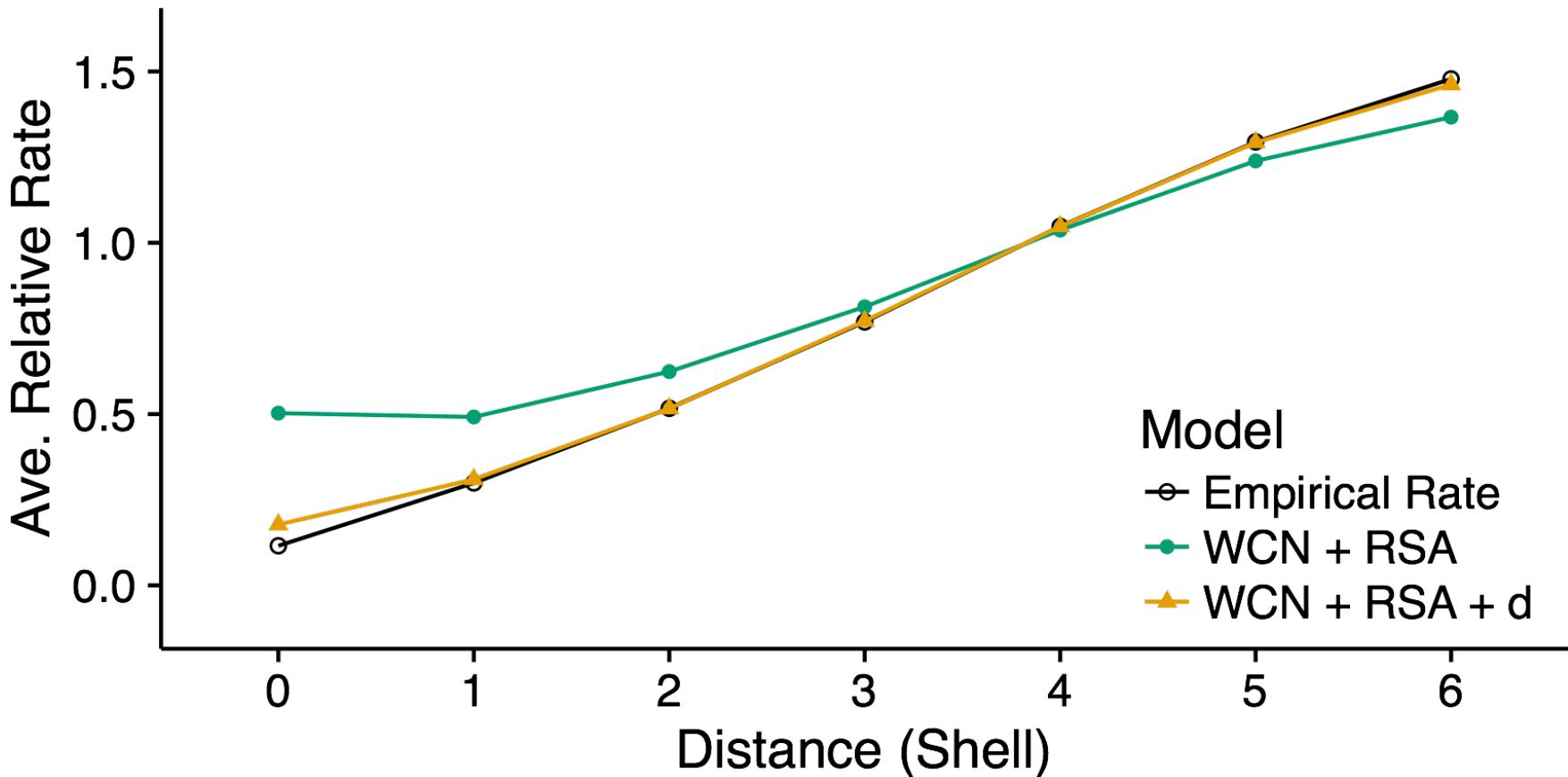
Distance is critical to accurately predict average rate near catalytic residues



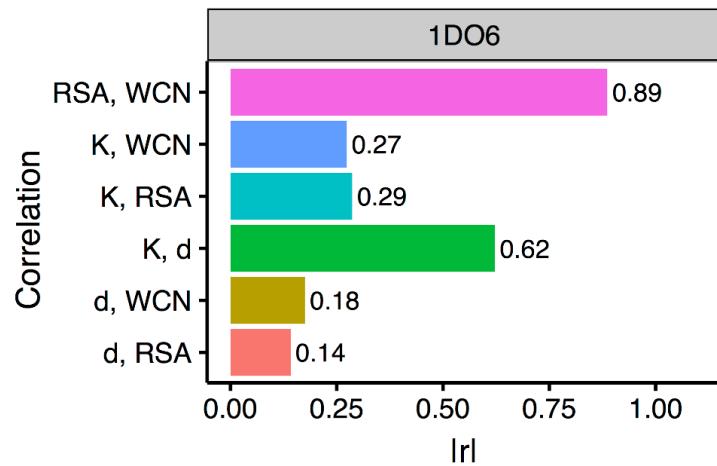
Distance is critical to accurately predict average rate near catalytic residues

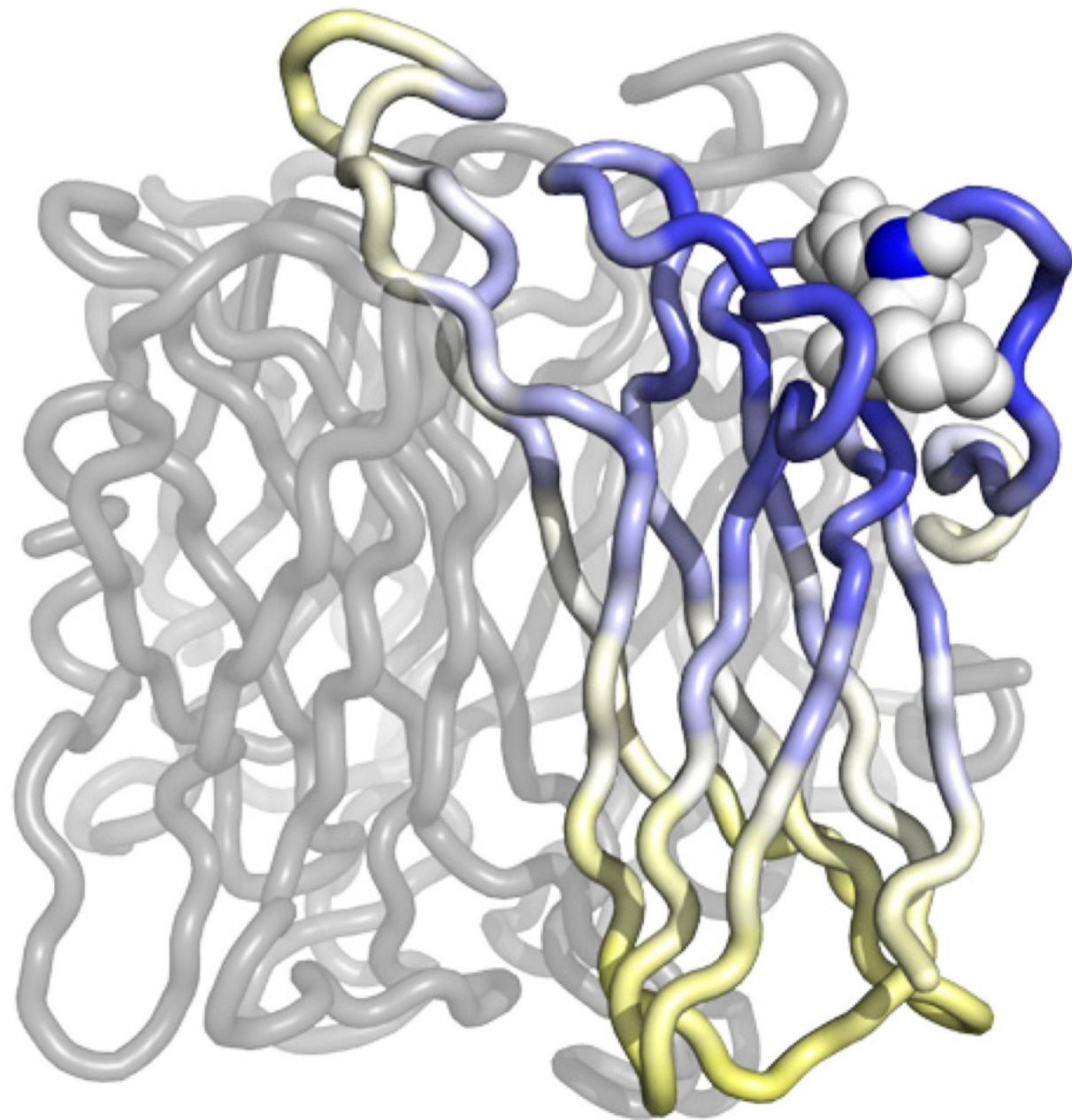


Distance is critical to accurately predict average rate near catalytic residues

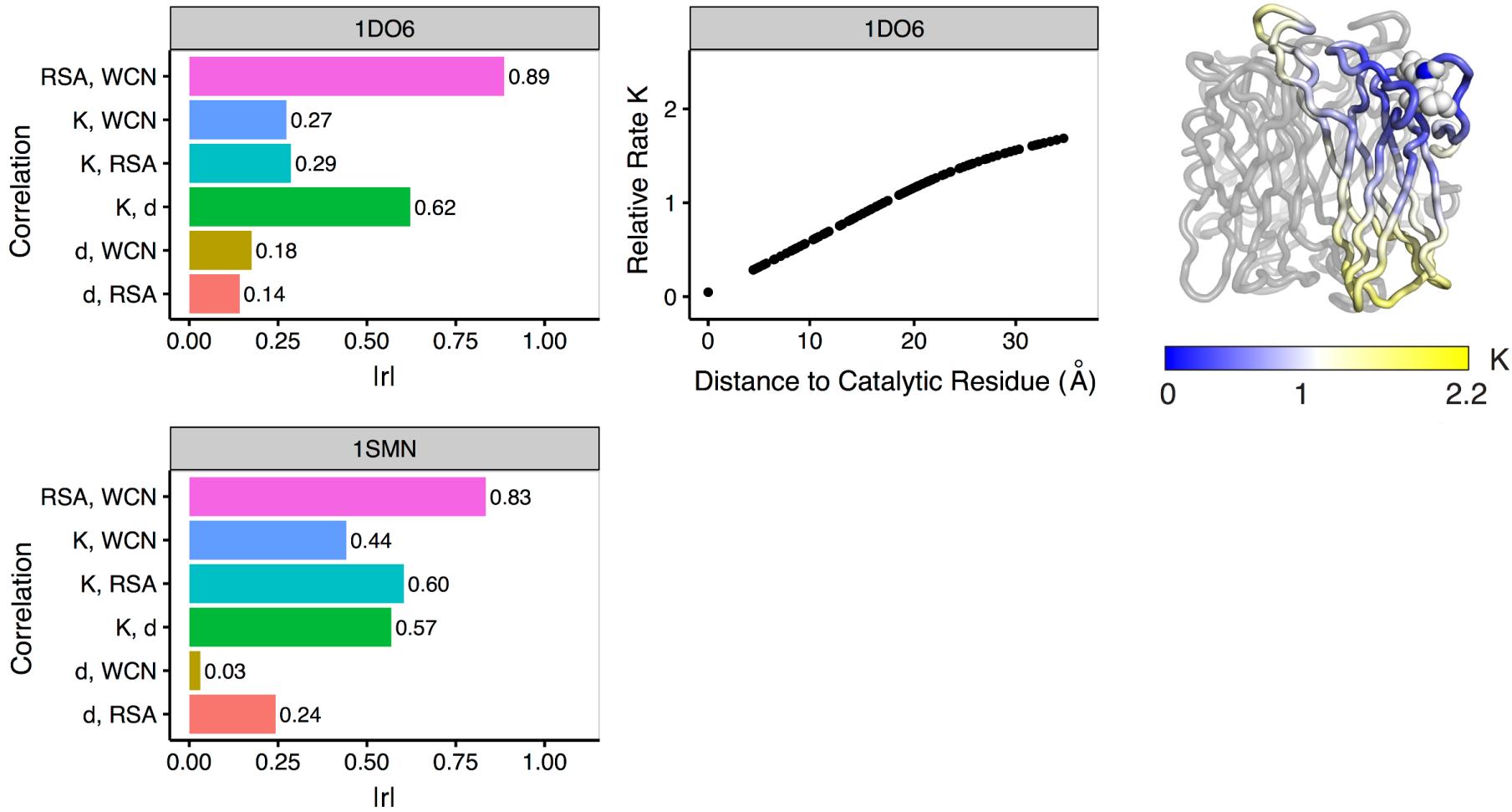


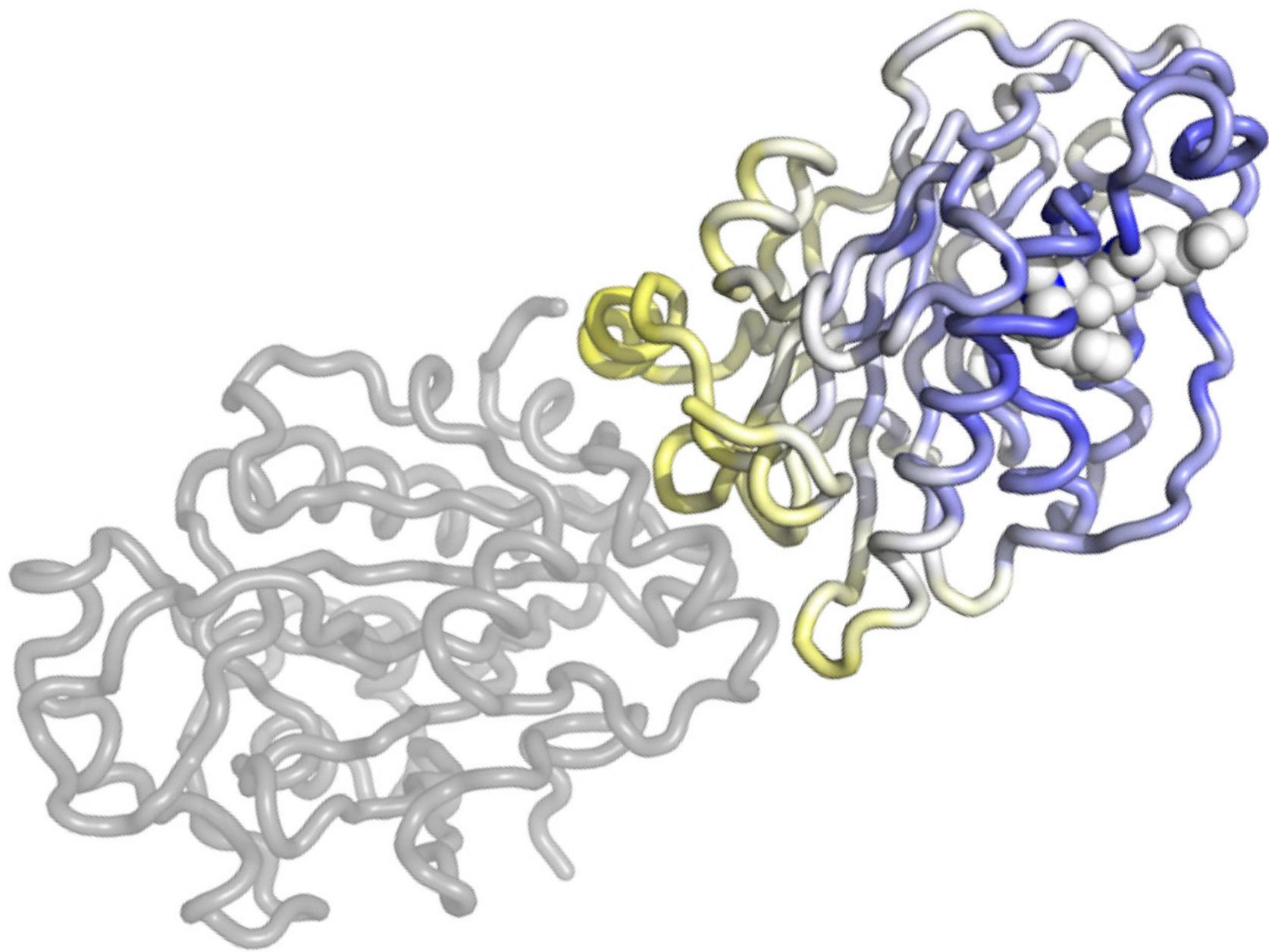
Effect is strongest for active sites on the protein surface



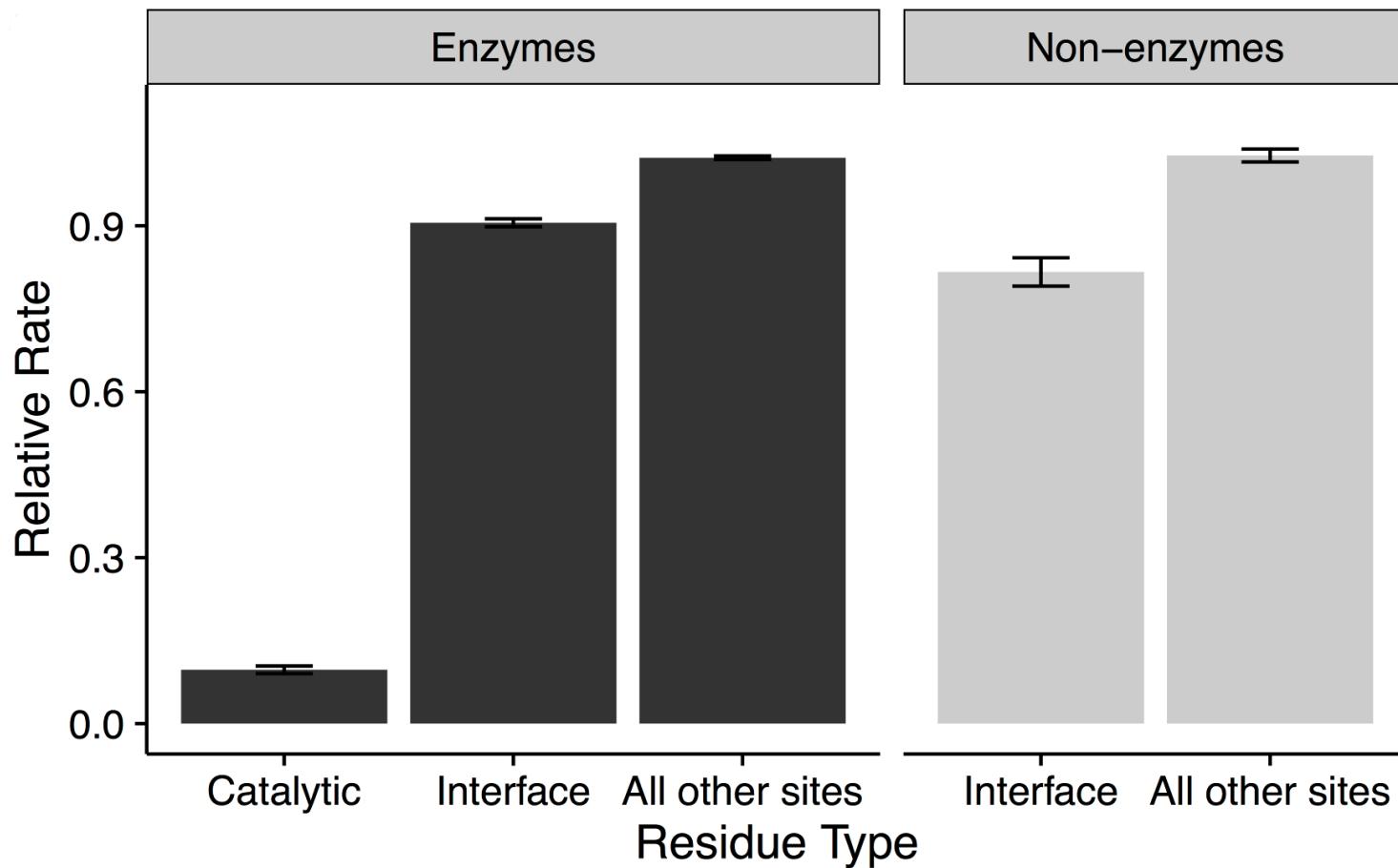


Effect is strongest for active sites on the protein surface





Catalytic residues impose much stronger constraints than interface residues



Take-home message

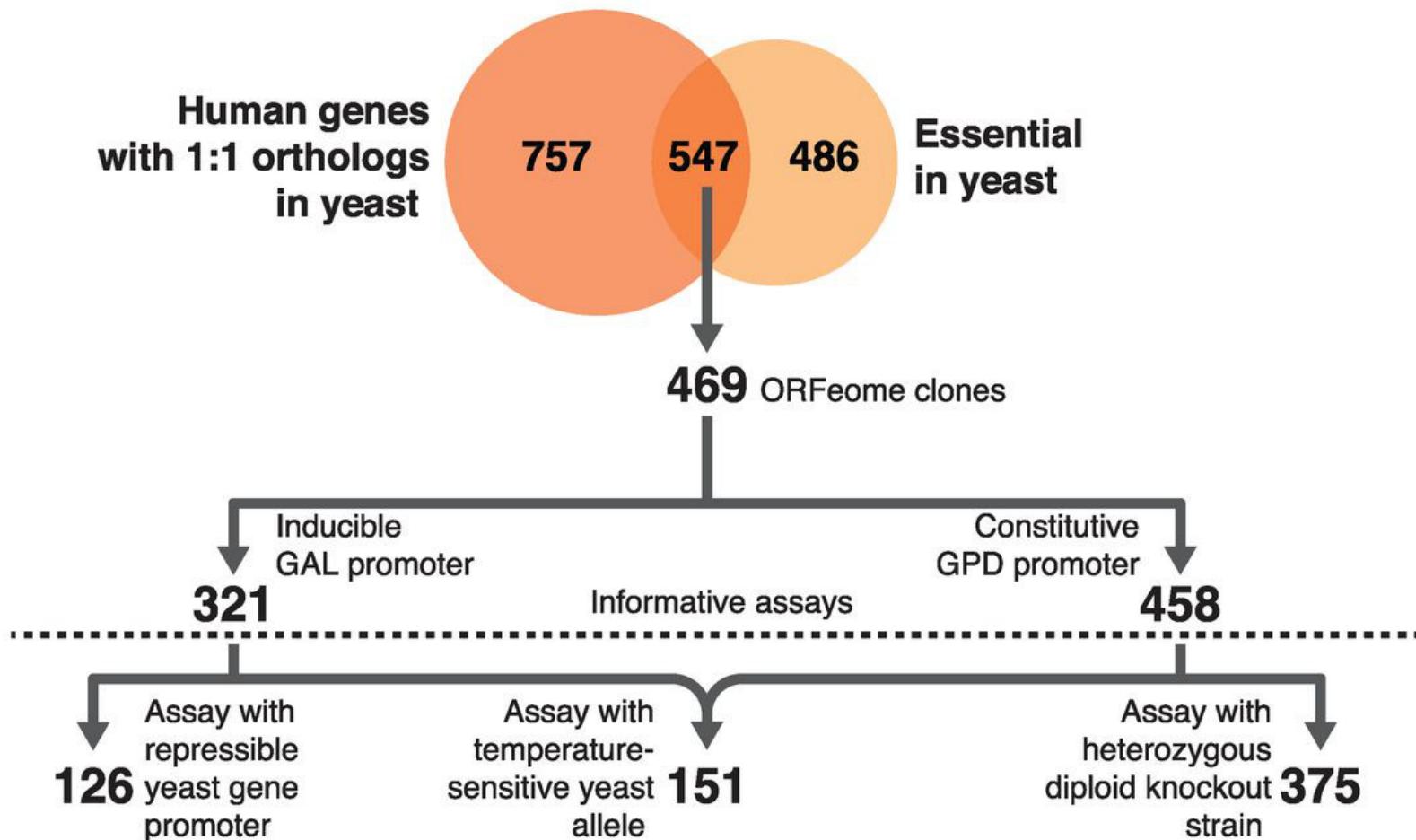
- Catalytic residues generate long-range selection gradients covering most of a typical enzyme structure

Part II: How does evolution constrain protein–protein interfaces?

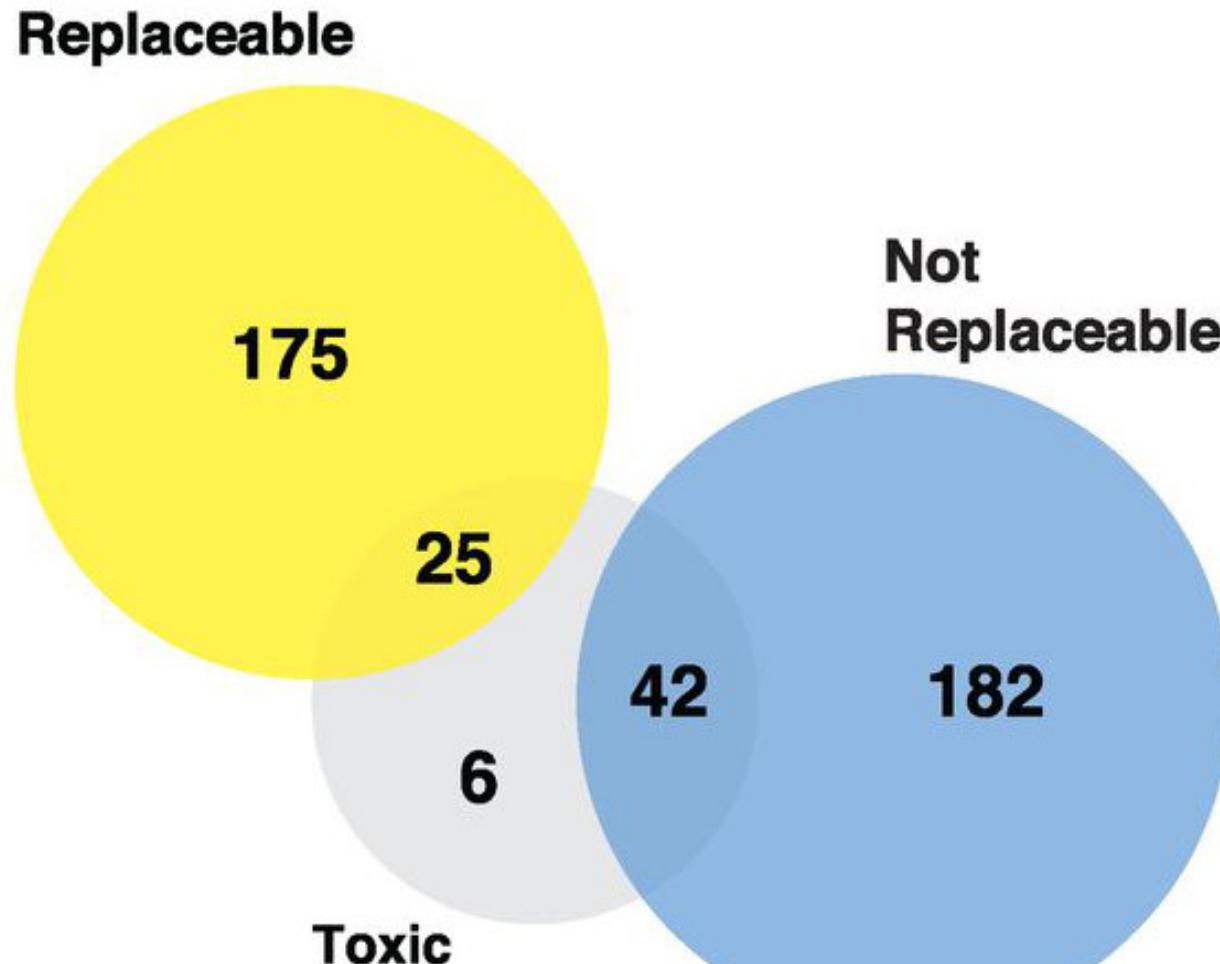
In collaboration with the Marcotte lab

Protein evolution simulations by Austin Meyer

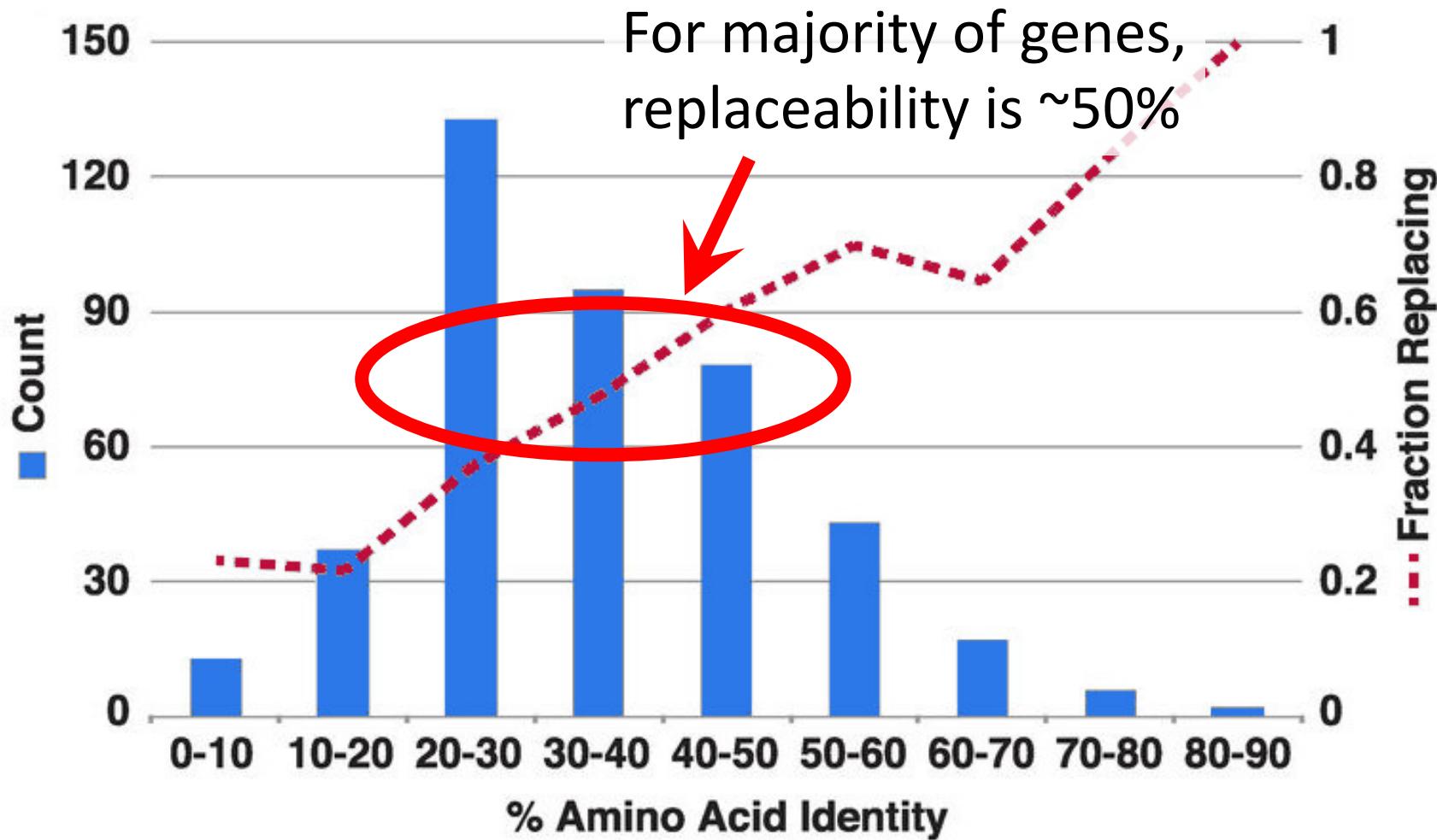
Can human genes function in a distant relative (baker's yeast)?



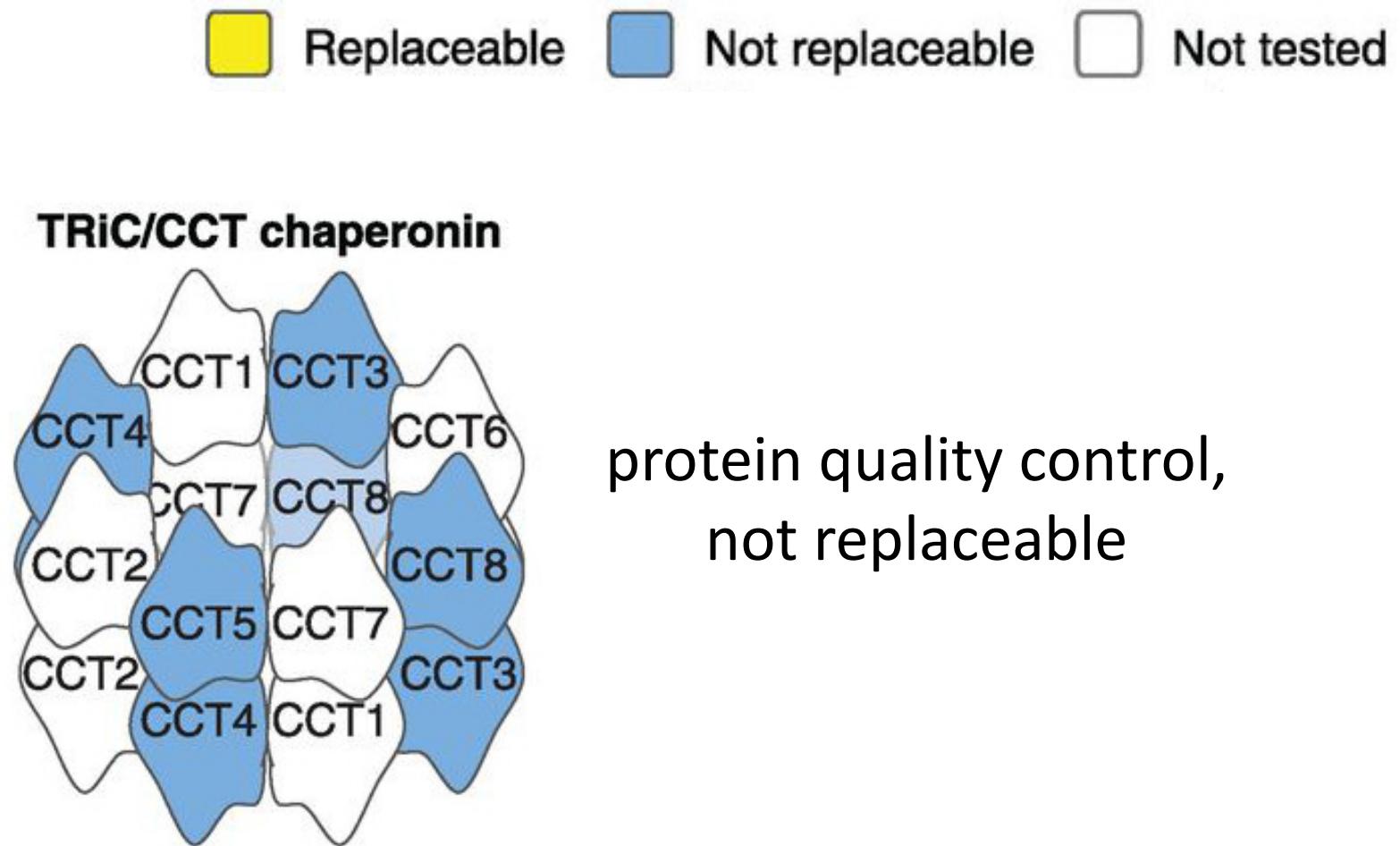
Yes!



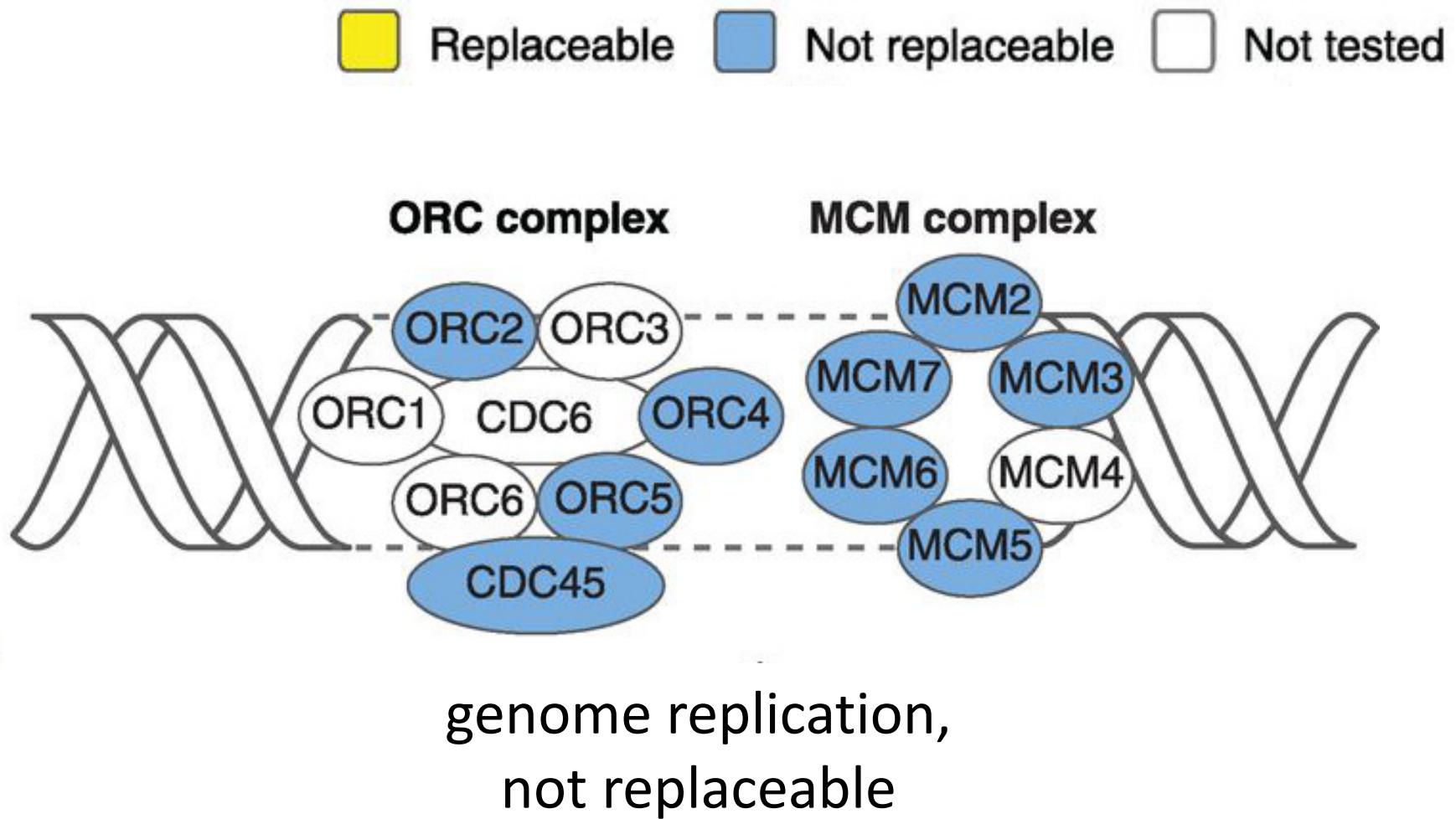
Sequence divergence is not a good predictor of replaceability



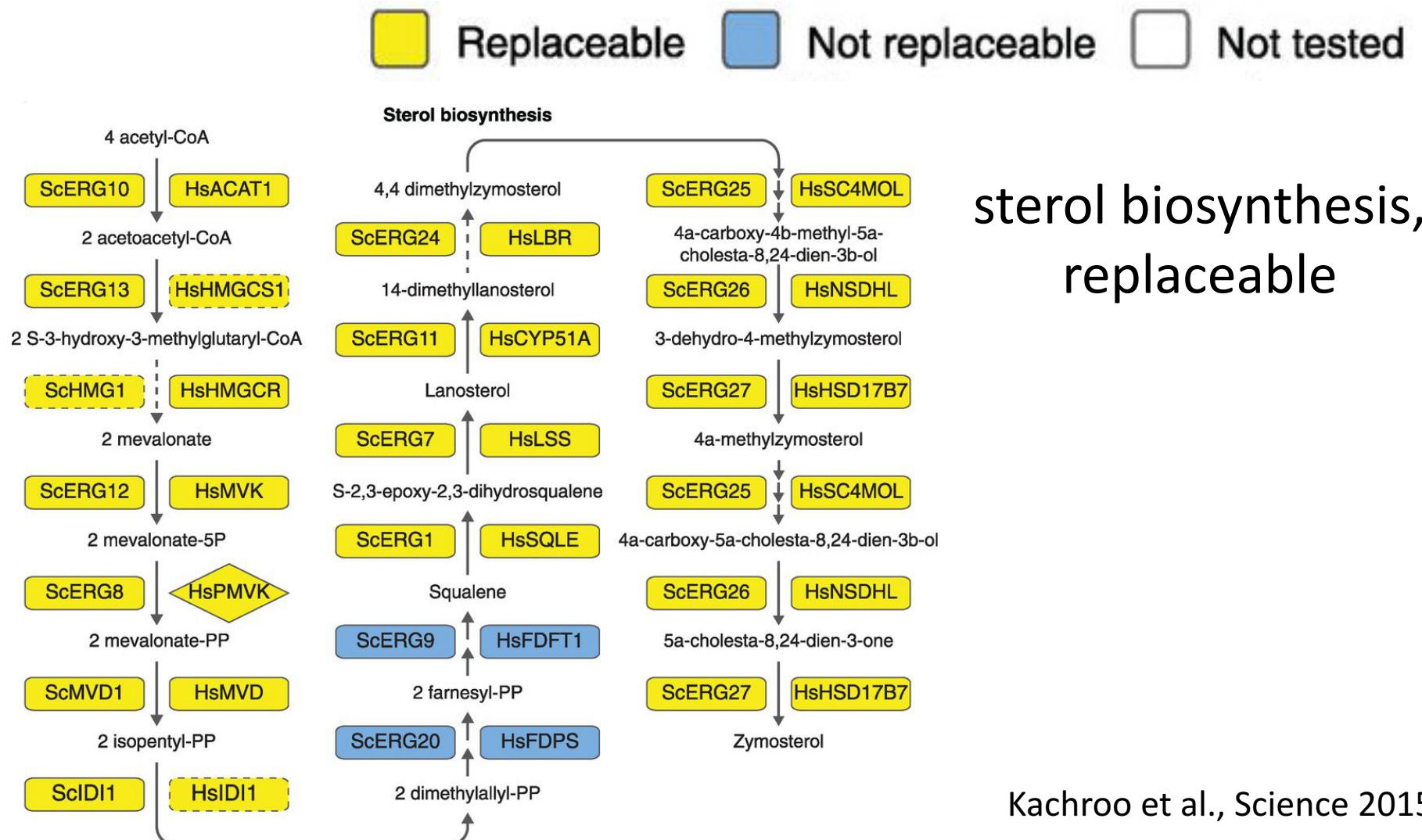
Genes tend to be replaceable in modules



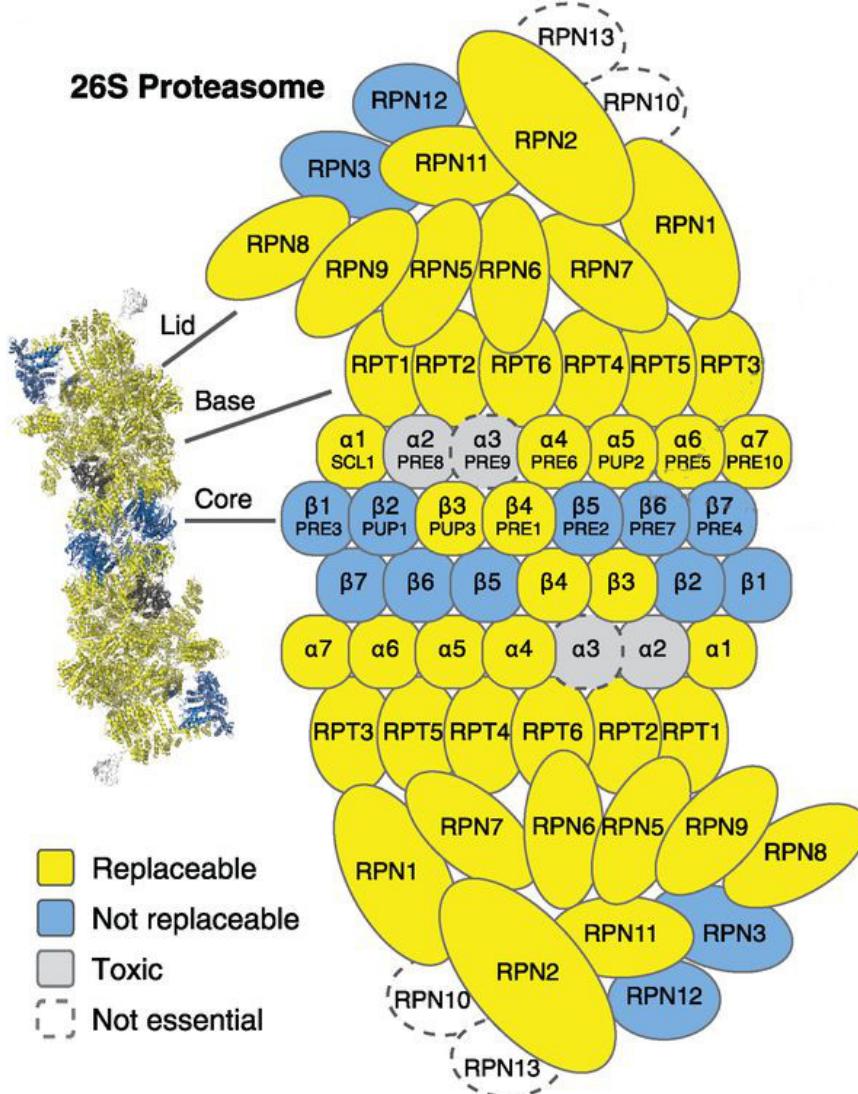
Genes tend to be replaceable in modules



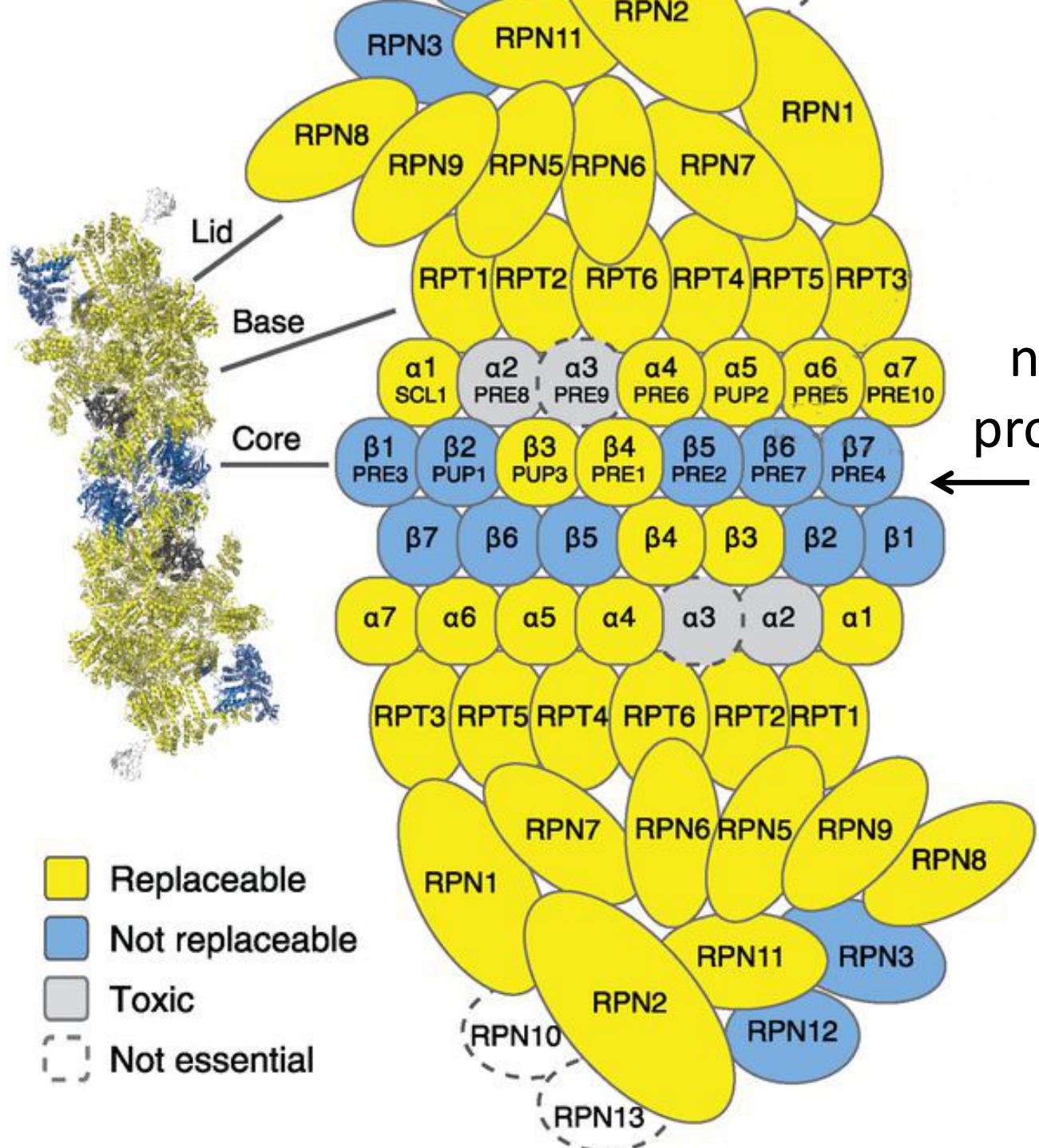
Genes tend to be replaceable in modules



Genes tend to be replaceable in modules

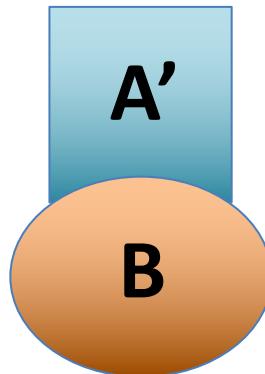
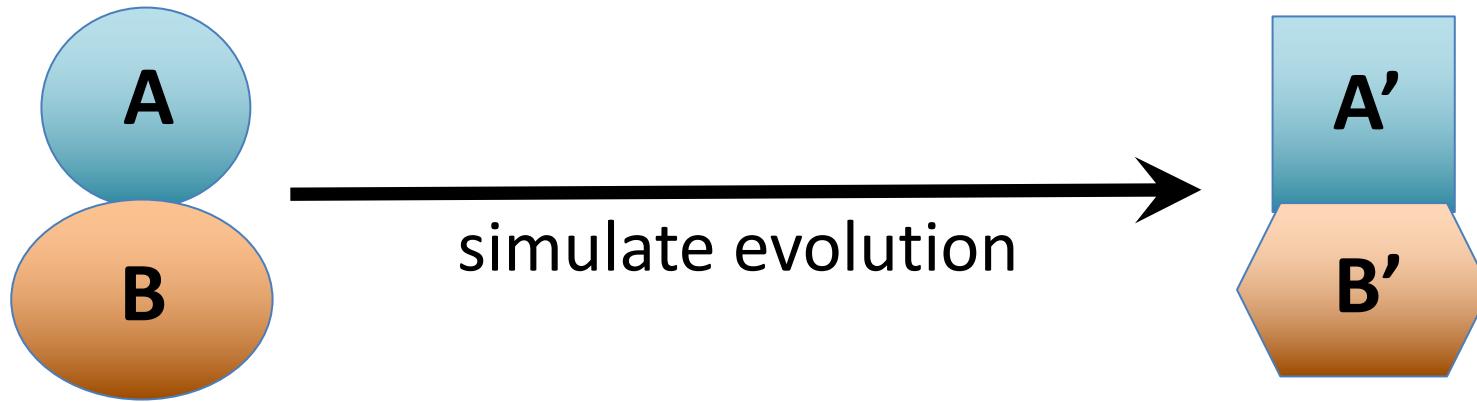


proteasome,
mostly replaceable



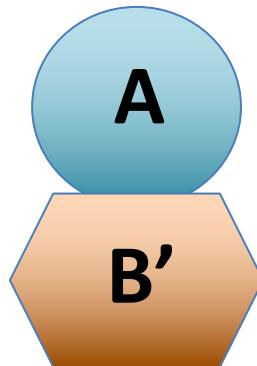
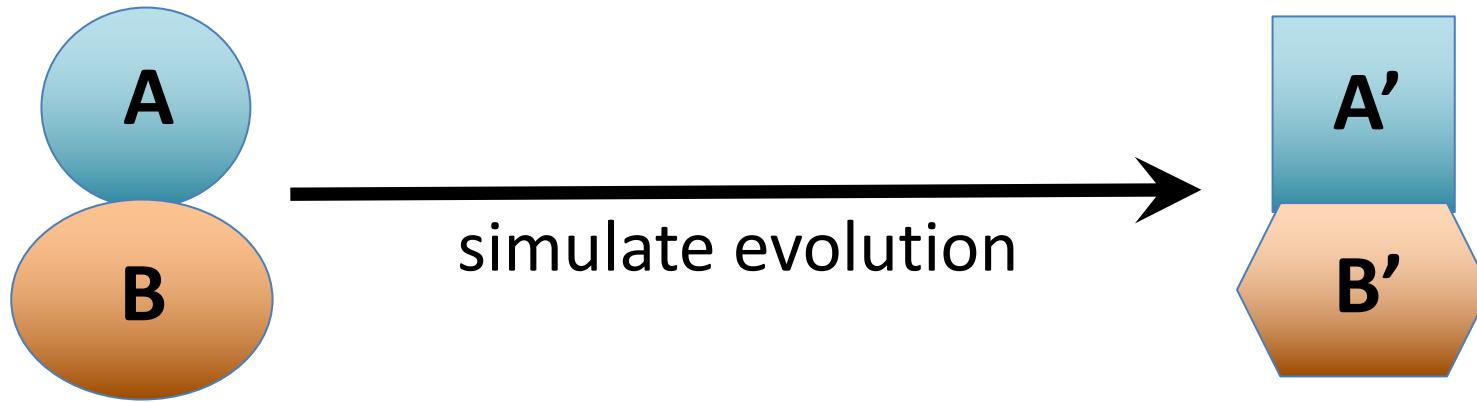
non-replaceable
proteins are mostly
in the β ring

Hypothesis: Conserved protein–protein interactions cause replaceability



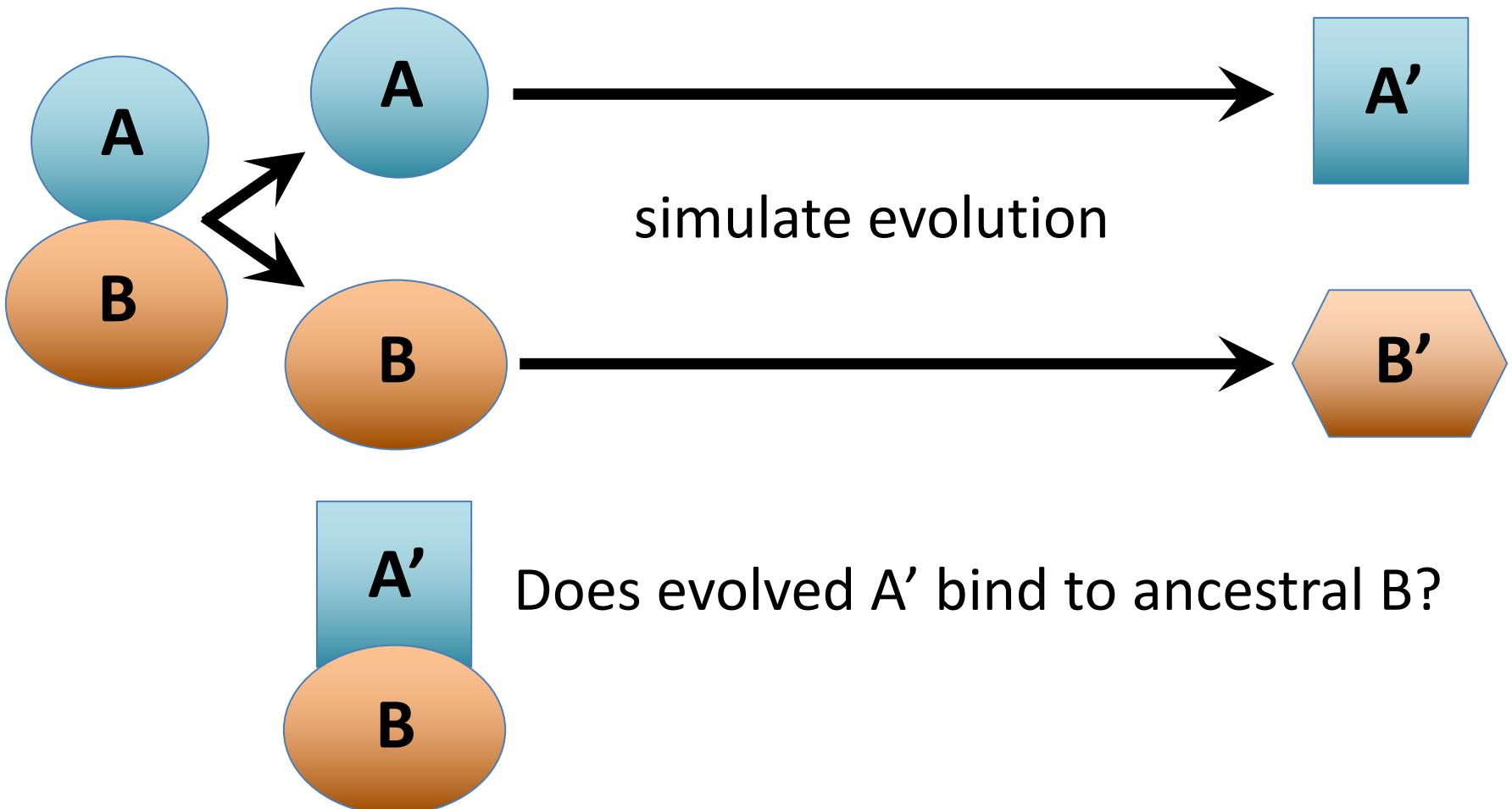
Does evolved A' bind to ancestral B?

Hypothesis: Conserved protein–protein interactions cause replaceability

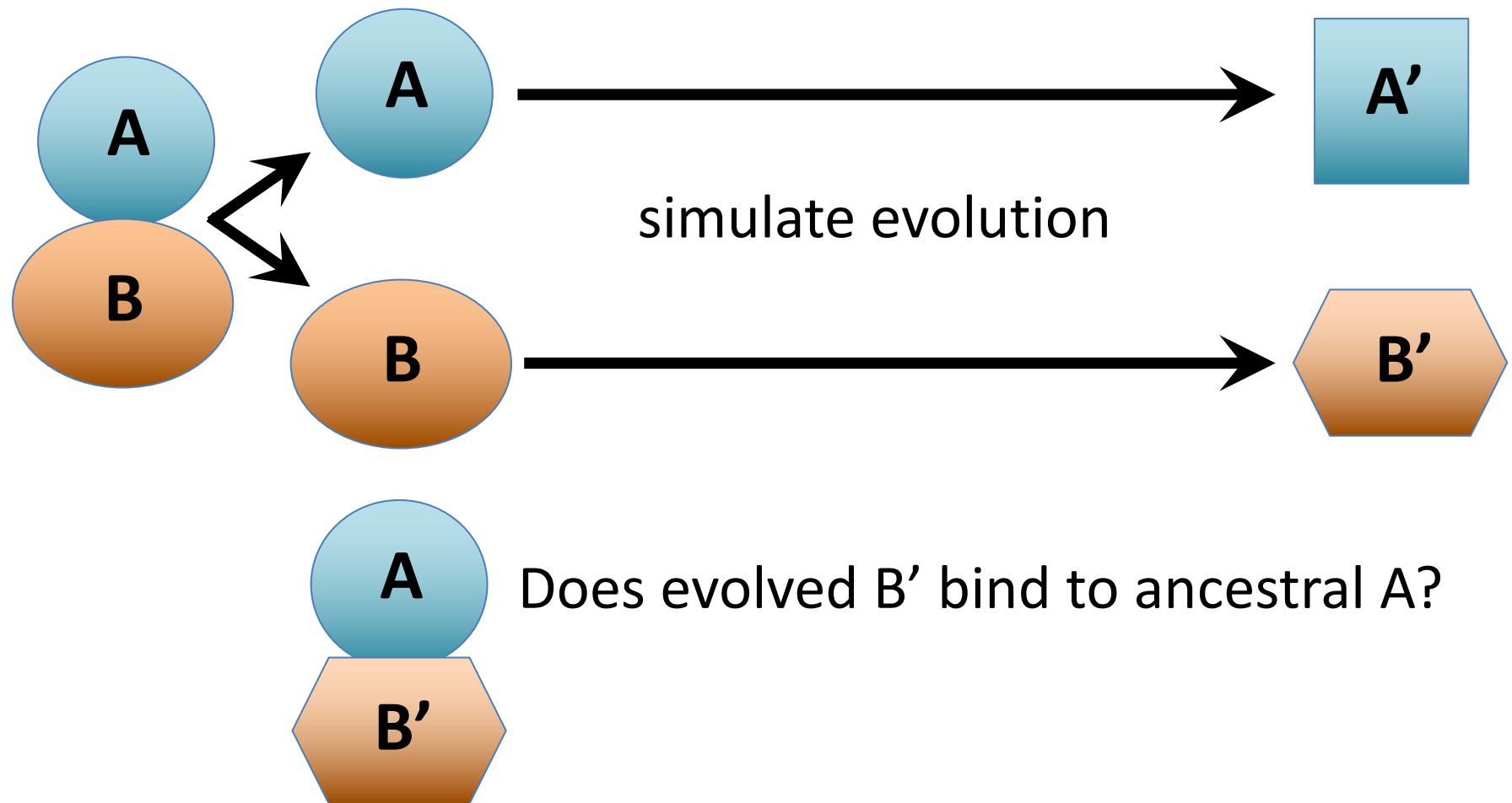


Does evolved B' bind to ancestral A?

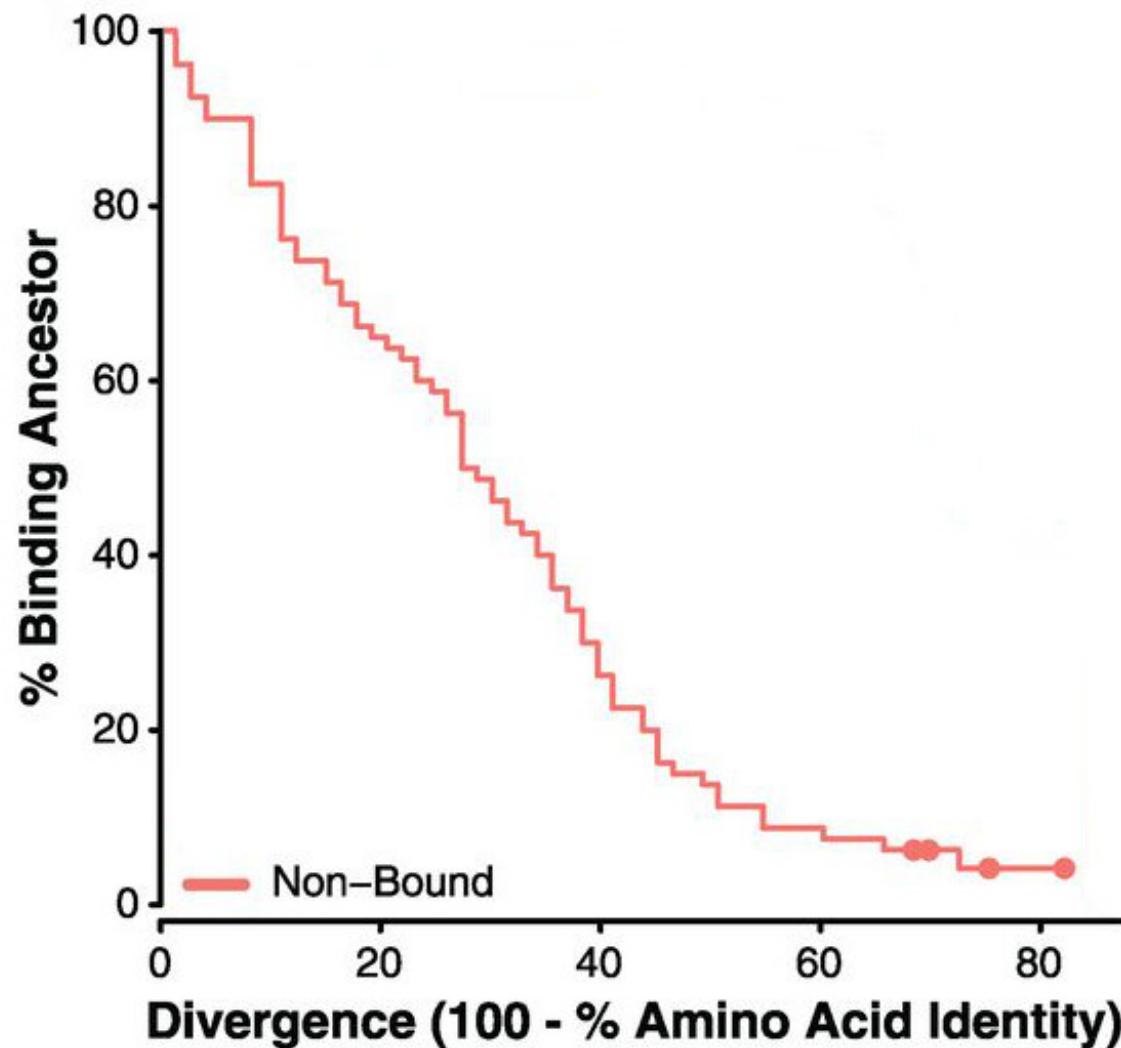
Control simulation: Evolve without binding



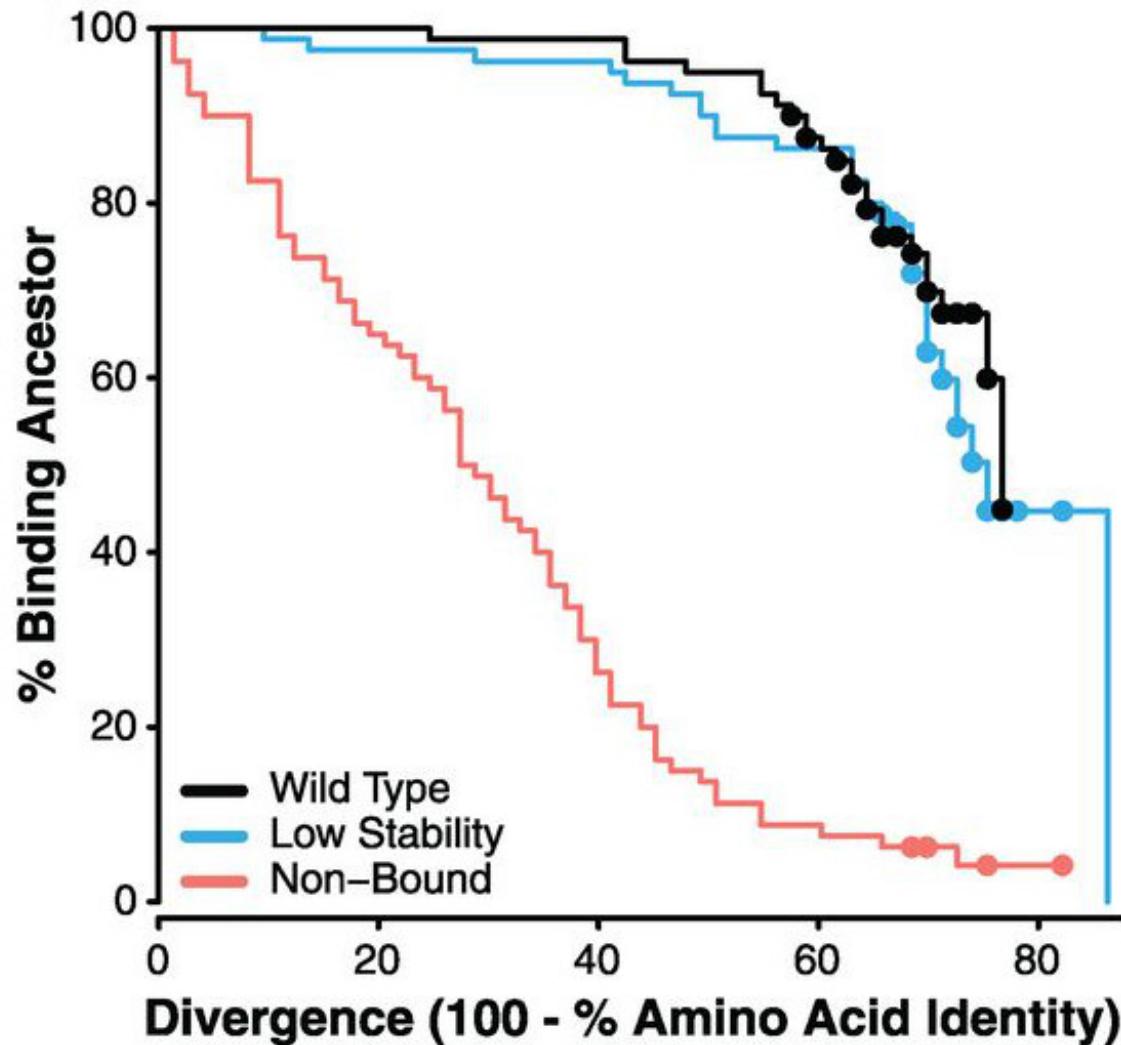
Control simulation: Evolve without binding



Binding to ancestor declines rapidly in the control scenario

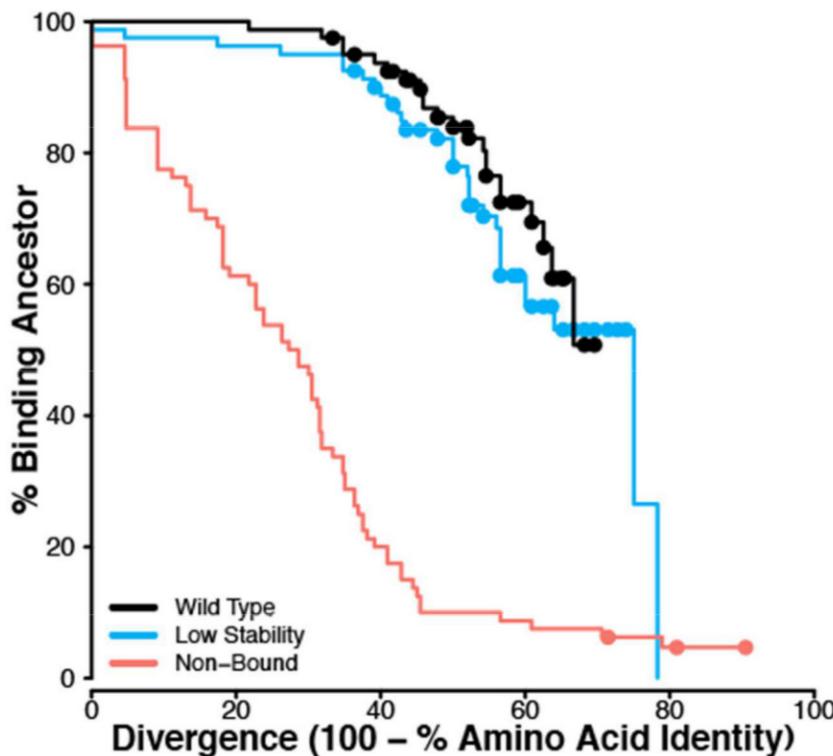


Binding to ancestor declines rapidly in the control scenario, but not in WT scenario

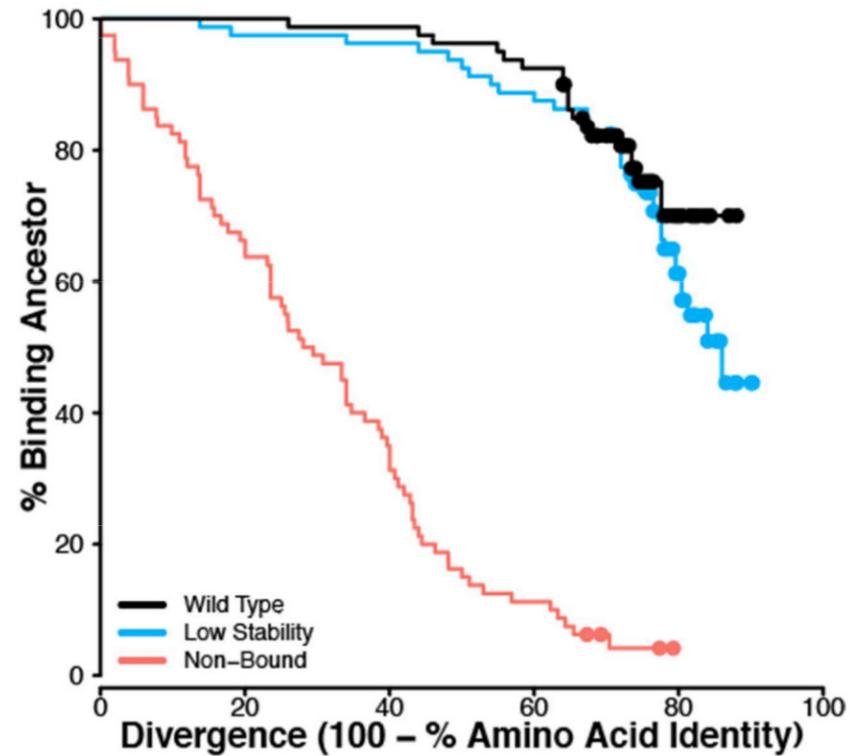


Even with substantial divergence in interface sites, evolved proteins bind ancestors

divergence in
interface sites



divergence in
non-interface sites



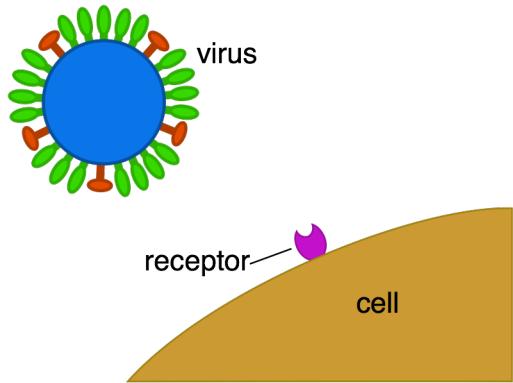
Take-home message

- As protein–protein interfaces diverge, they seem to maintain biochemical similarity to their ancestral state

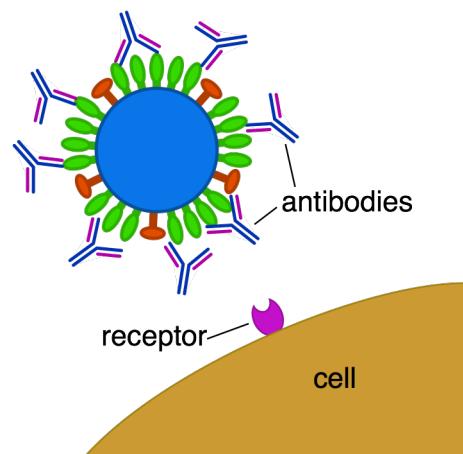
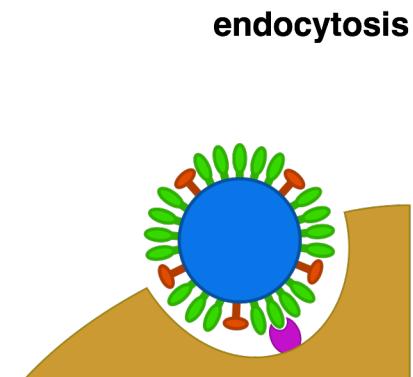
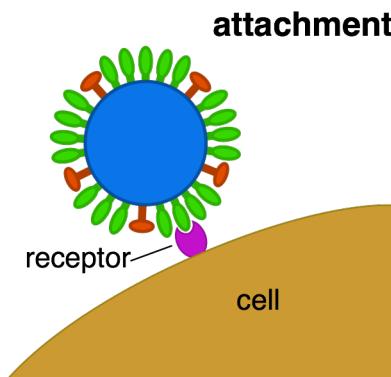
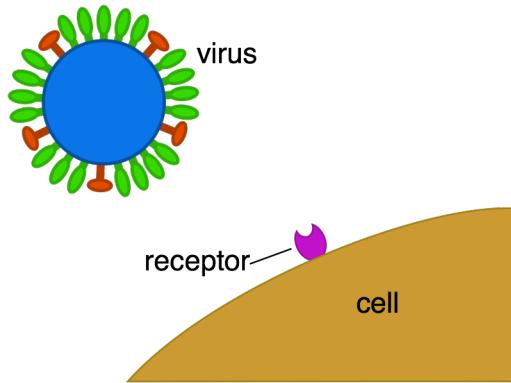
Part III: Can we identify sites under positive selection in influenza HA?

Work by Claire McWhite (Marcotte lab), Austin Meyer

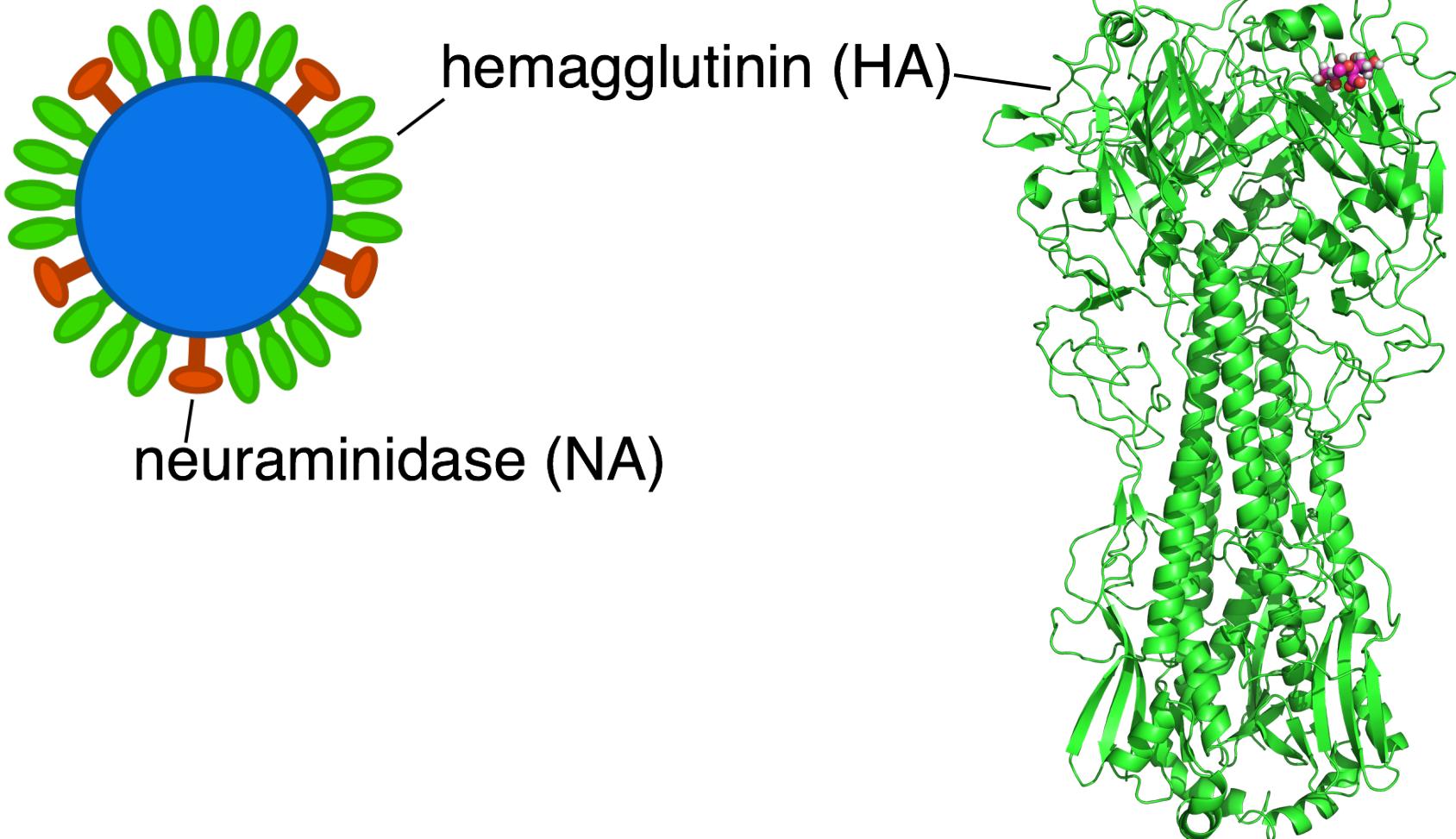
Influenza enters cells via attachment and endocytosis



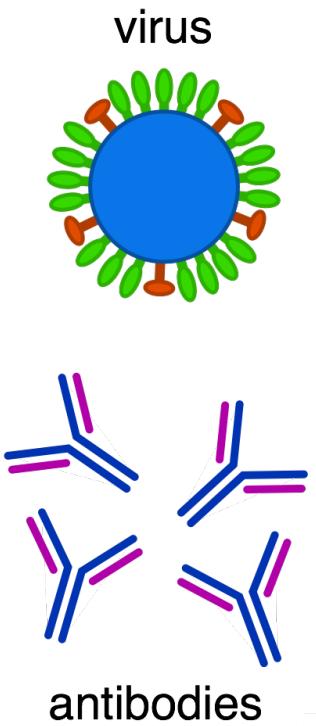
Antibodies bind to viral surface proteins and prevent attachment



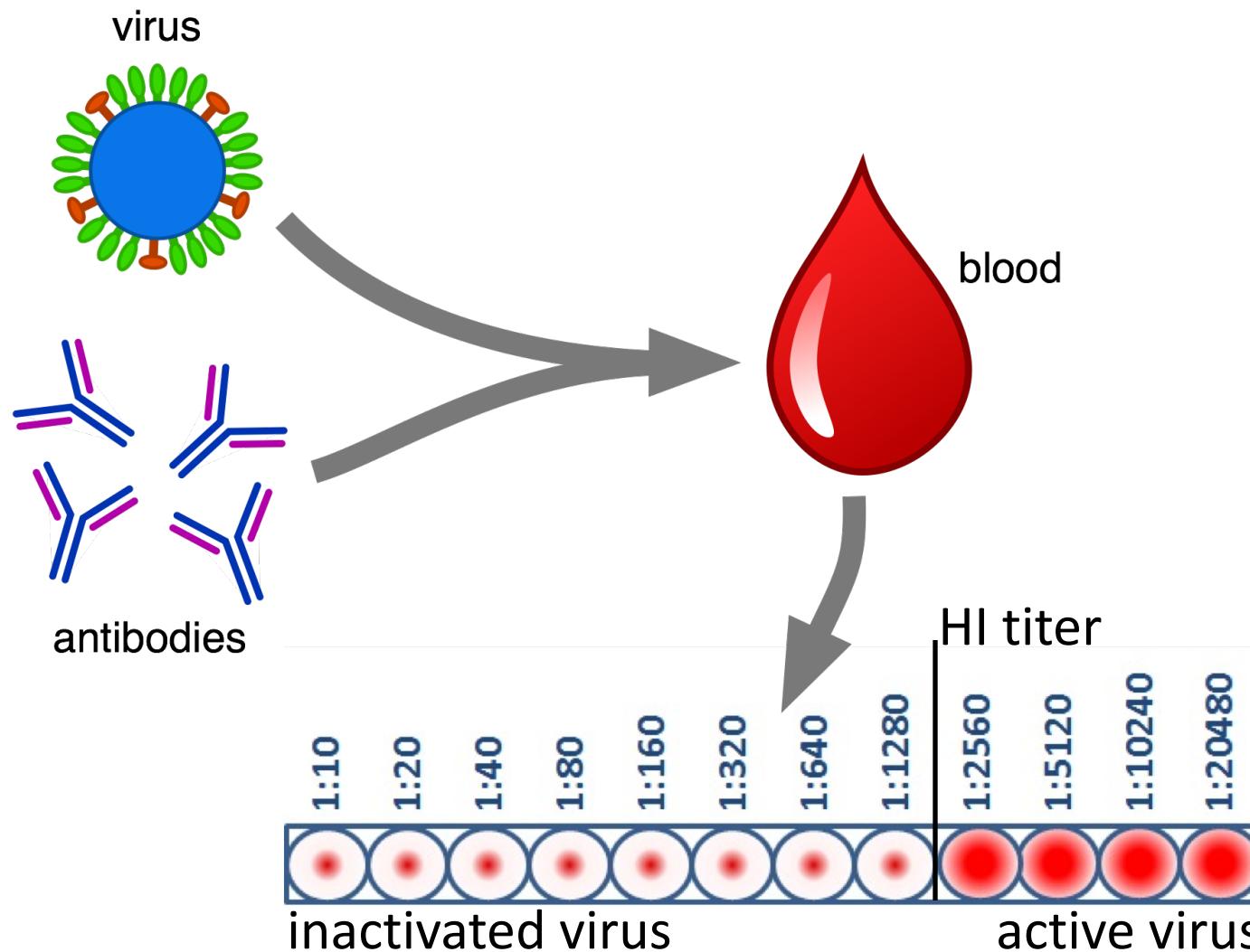
The primary antigenic target is the hemagglutinin protein (HA)



The hemagglutinin inhibition assay characterizes a virus's antigenic profile

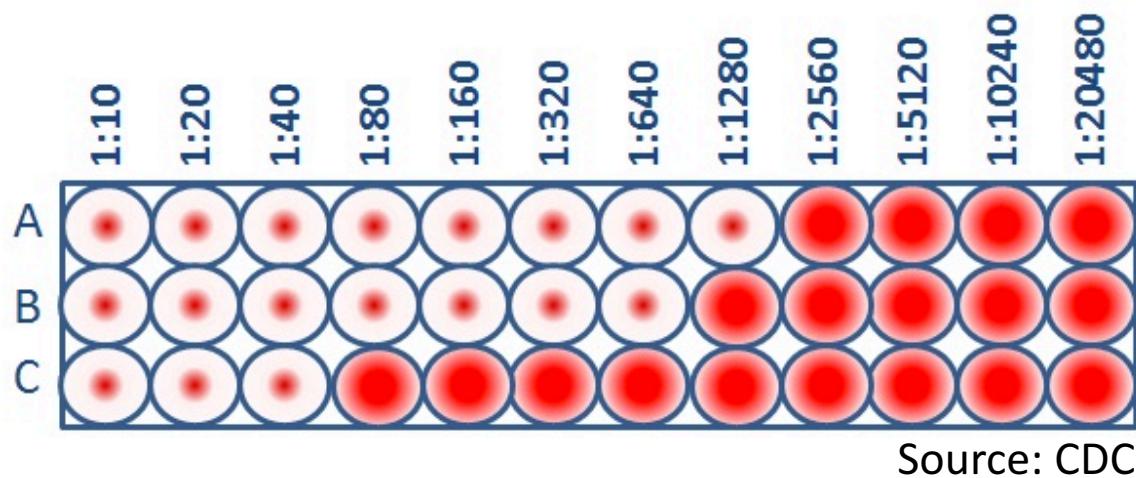


The hemagglutinin inhibition assay characterizes a virus's antigenic profile

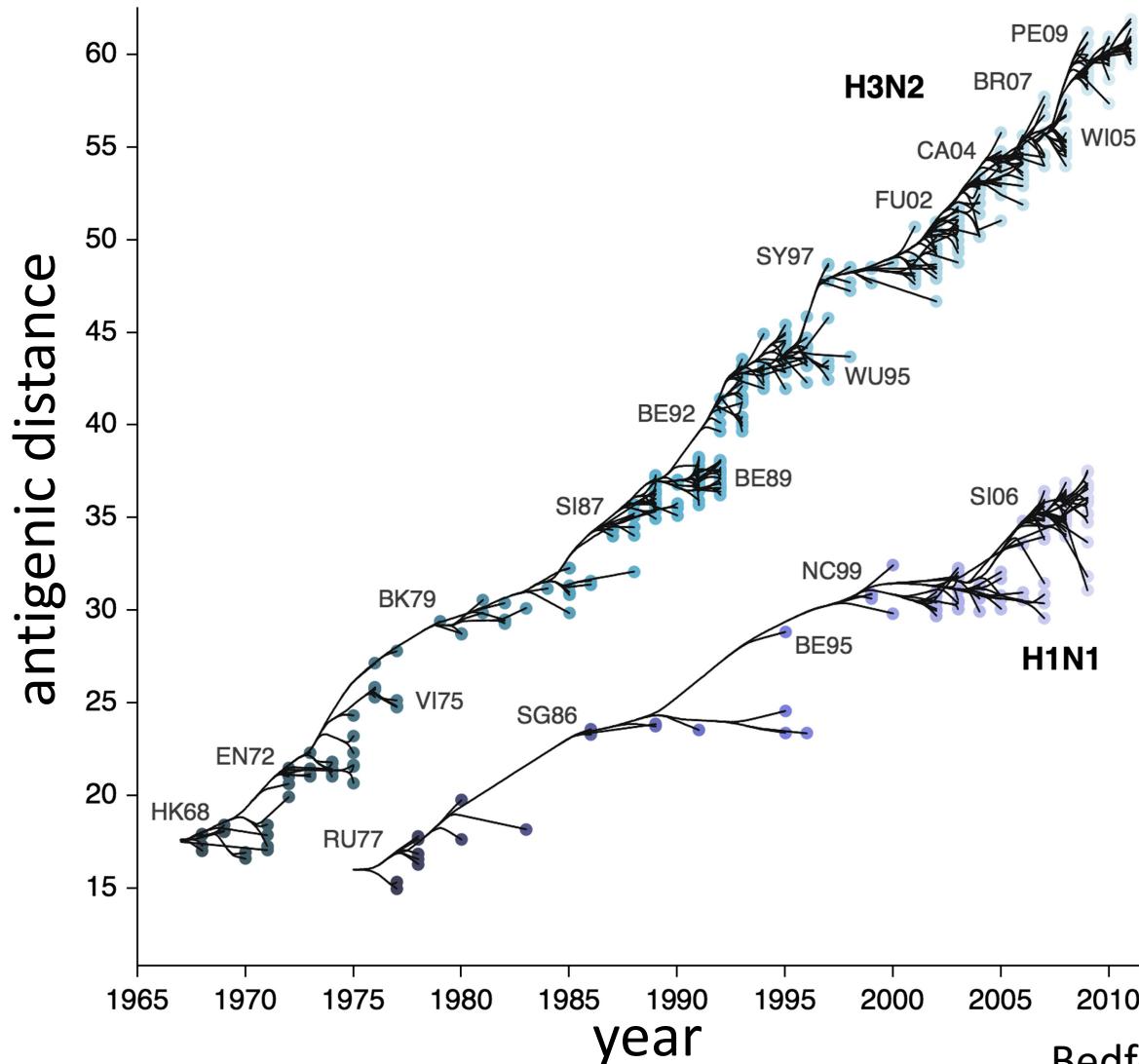


The hemagglutinin inhibition assay characterizes a virus's antigenic profile

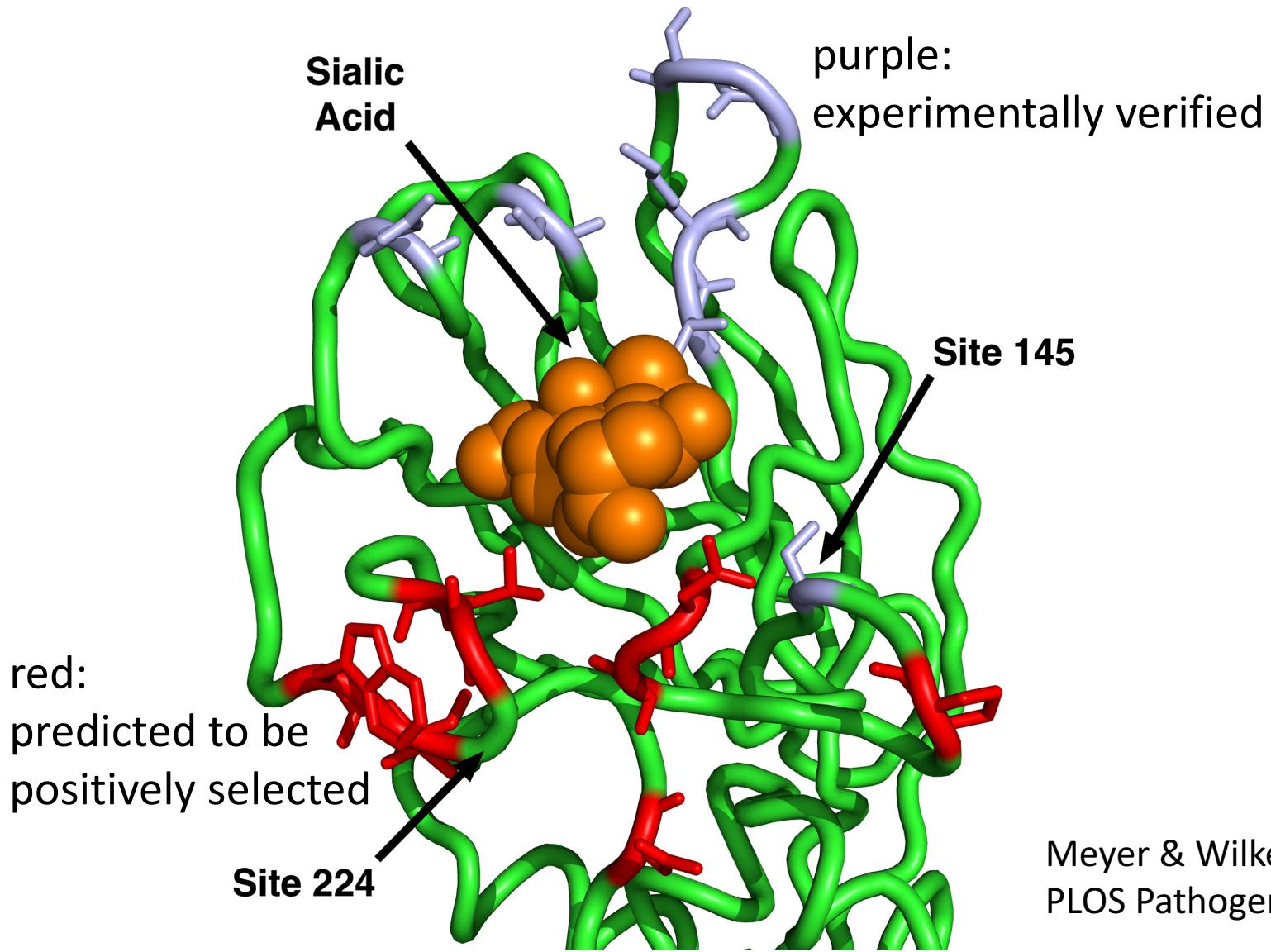
Previous Season's Vaccine Virus
Circulating Virus 1 ("like" virus)
Circulating Virus 2 (low reactor)



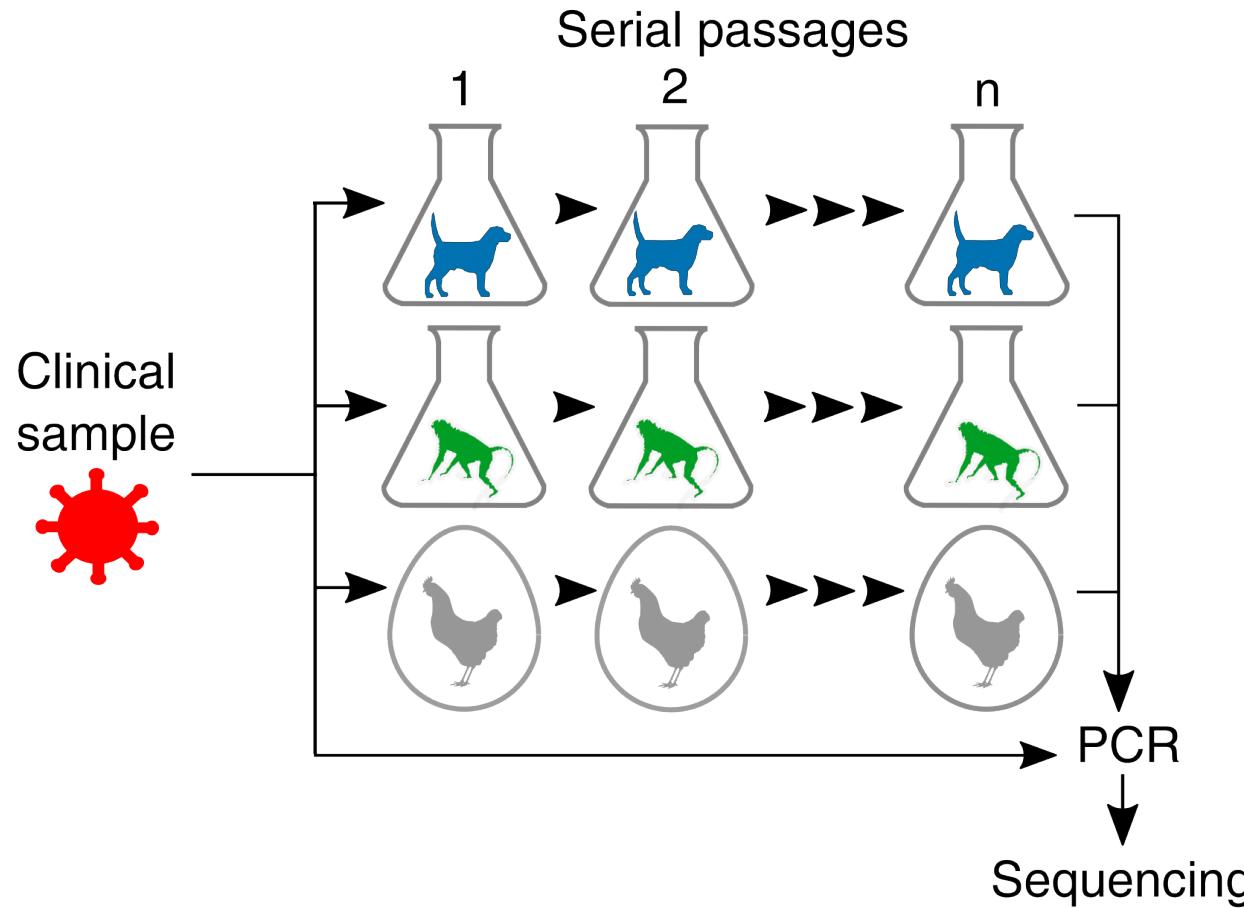
Influenza viruses rapidly evolve in response to the human immune system



Positively selected sites disagree with experimentally verified antigenic sites

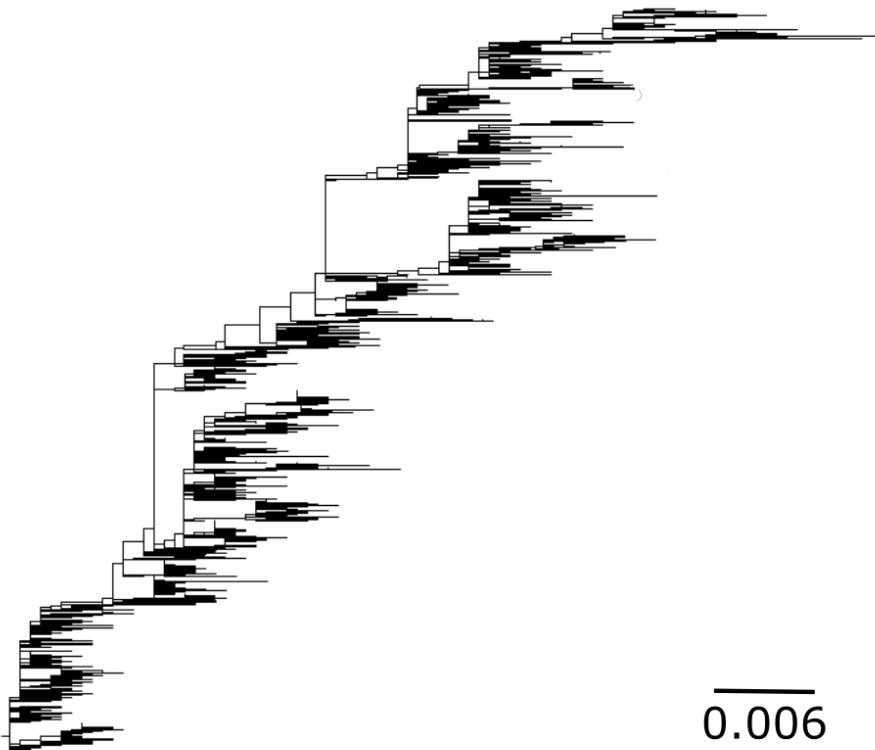


Influenza virus is commonly passaged before analysis, to obtain high viral titers



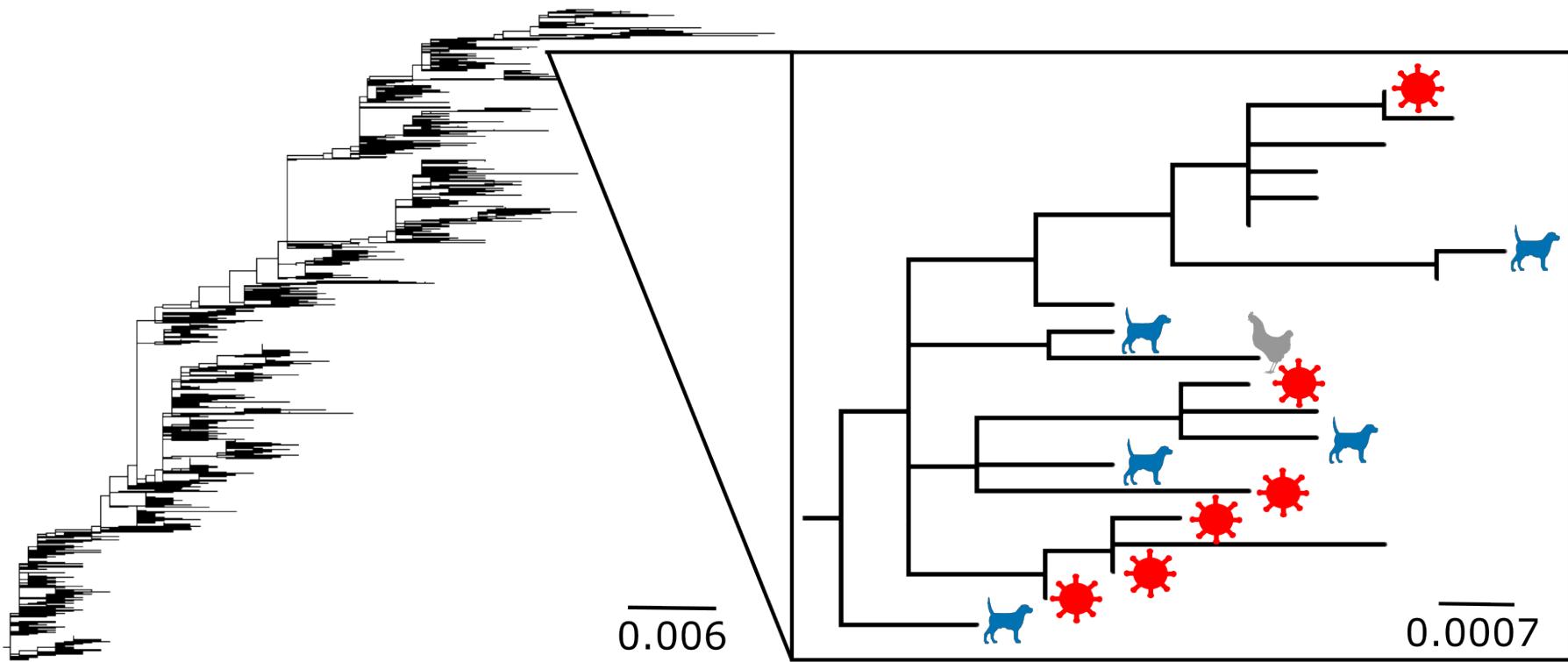
Passaging treatments are intermingled in the phylogeny

Influenza H3N2 phylogeny, 2005–2015



Passaging treatments are intermingled in the phylogeny

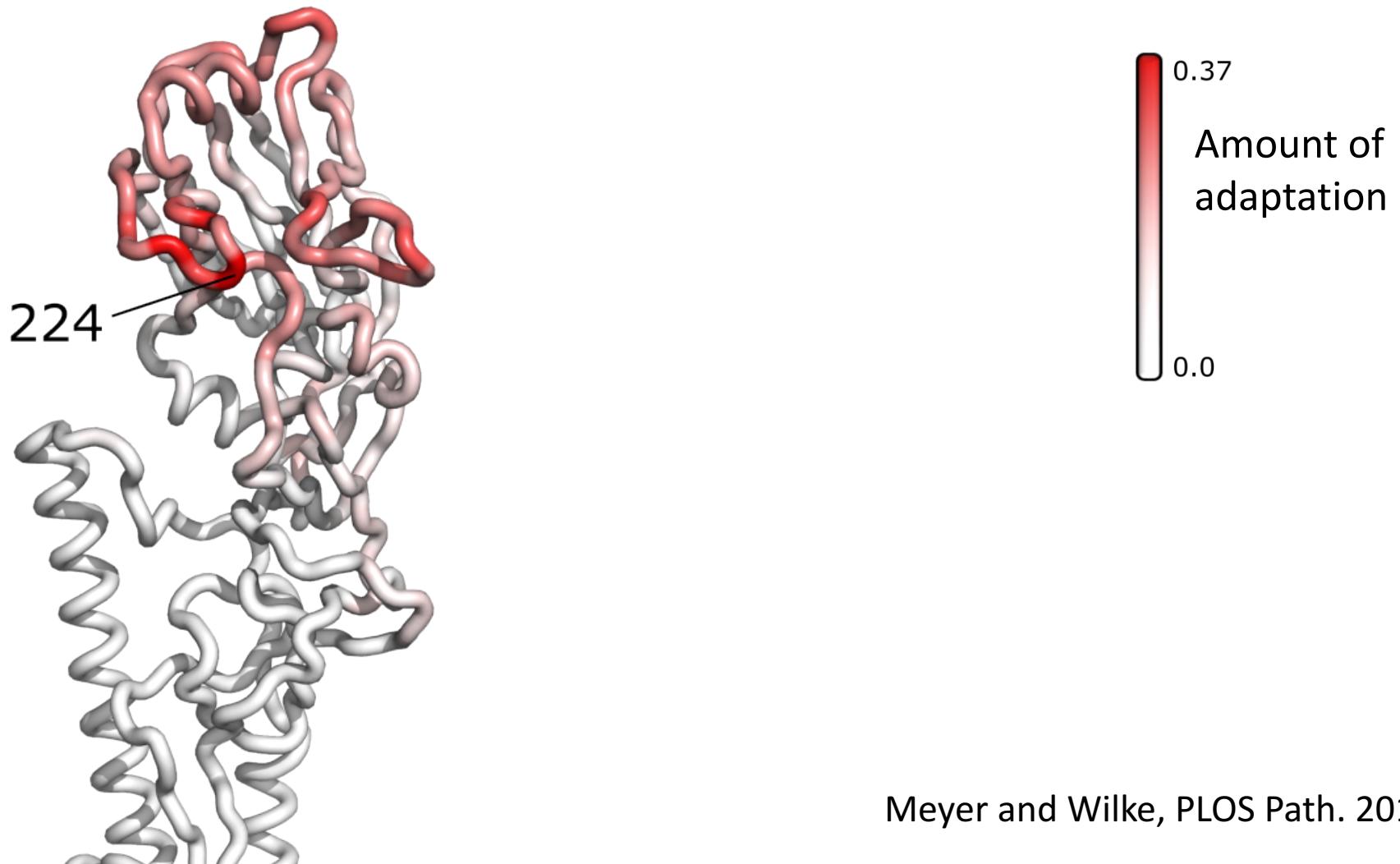
Influenza H3N2 phylogeny, 2005–2015



Unfortunately, passaging annotations are a mess

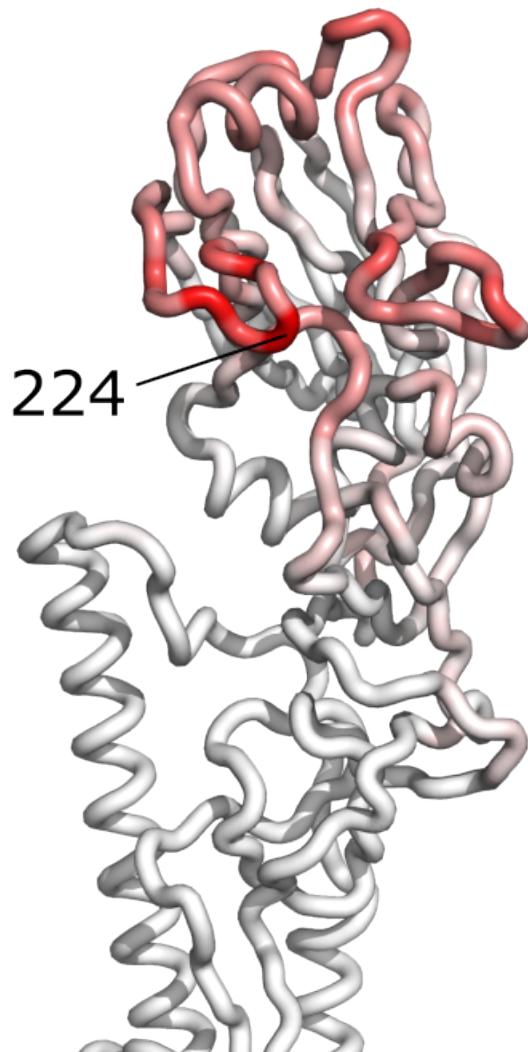
cell	egg	monkey	unpassaged
XMDCK2	E5_E1	PRHMK_2	ORIGINAL
C_3_1	AM1	R_MIX_1_RHMK_1	LUNG_1
MX_S2	E4_E1	PRHMK_3	ISOLATED_DIRECTLY_FRO
CX_S2	EX_E1	RMK_1ST_PASSAGE	M_HOST_NO_PASSA
C_2_1	SPFCK2E3	RHMK_5	GE
MDCK_1	PASSAGE_DETAILS__CK3_	RMK1	P0
M_1_C_2_SIAT1	_E6	RHMK1	OR
M1M2_C1	E6	R1	
MDCK1_SIAT1_SIAT1	E4E8_E1	RII	
PASSAGE_DETAILS__ND_ MDCK	E4_E2_E1 SPFCK3E3	PRHMK_1 PASSAGE_DETAILS__PMK	
S3	EGG	01	
X1_S2	EGG_PASSAGE	RHMK_1	
...	

Positive selection in H₃N₂ HA near site 224

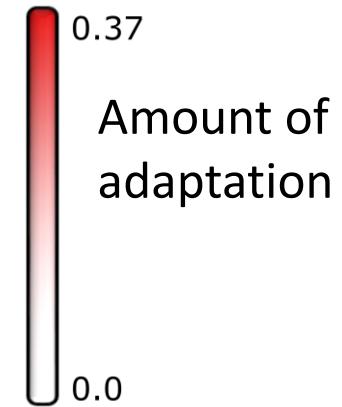
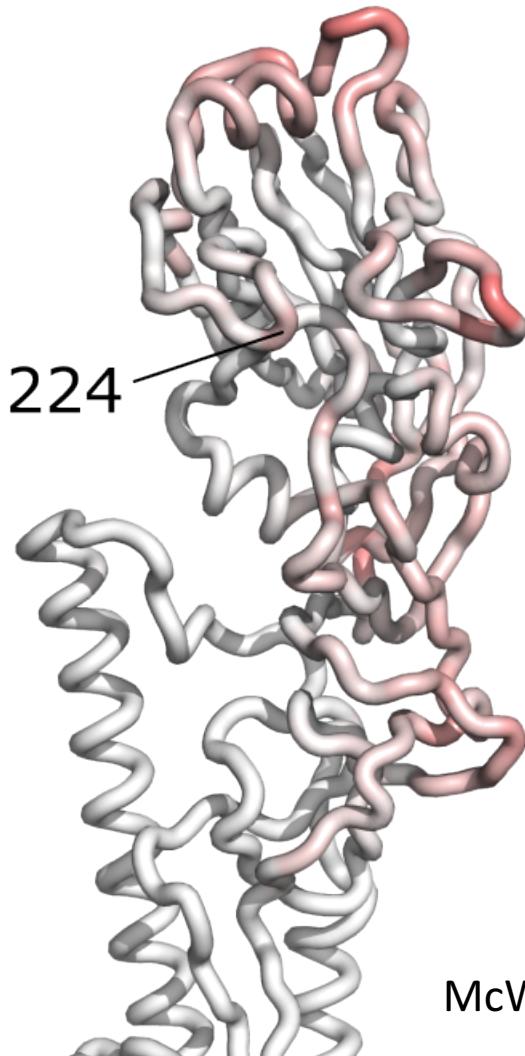


Positive selection in H₃N₂ HA near site 224 is caused by adaptation to MDCK cells

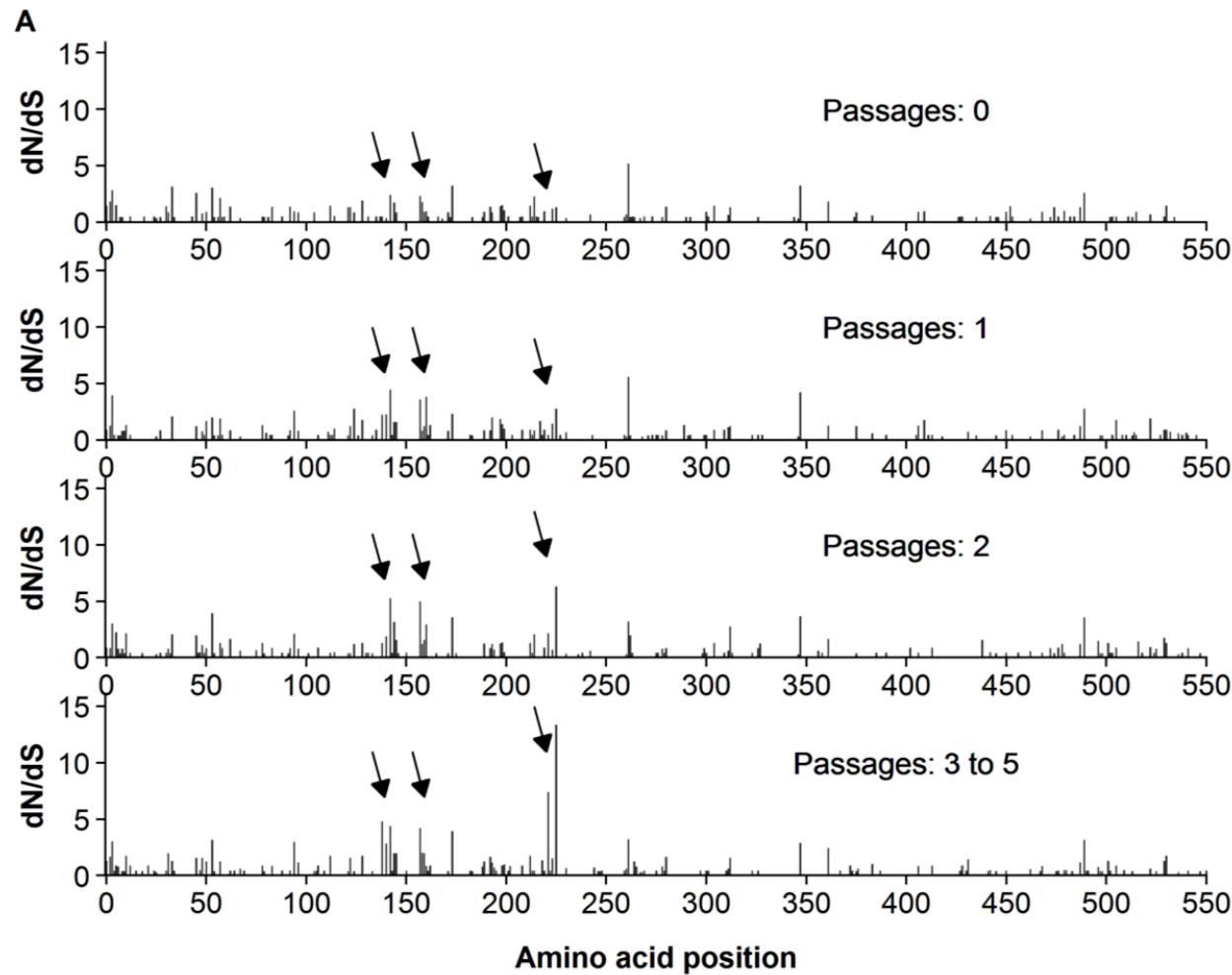
MDCK passaged virus



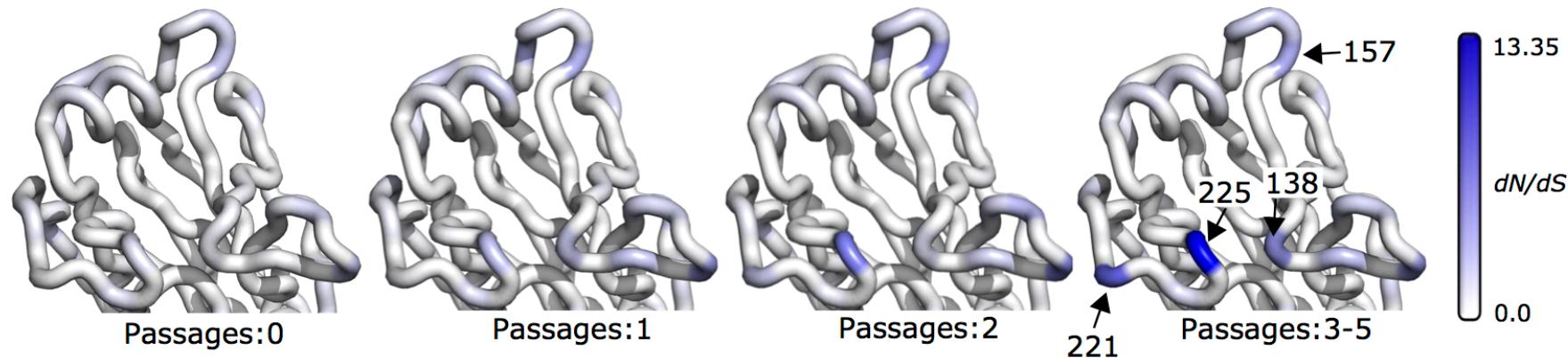
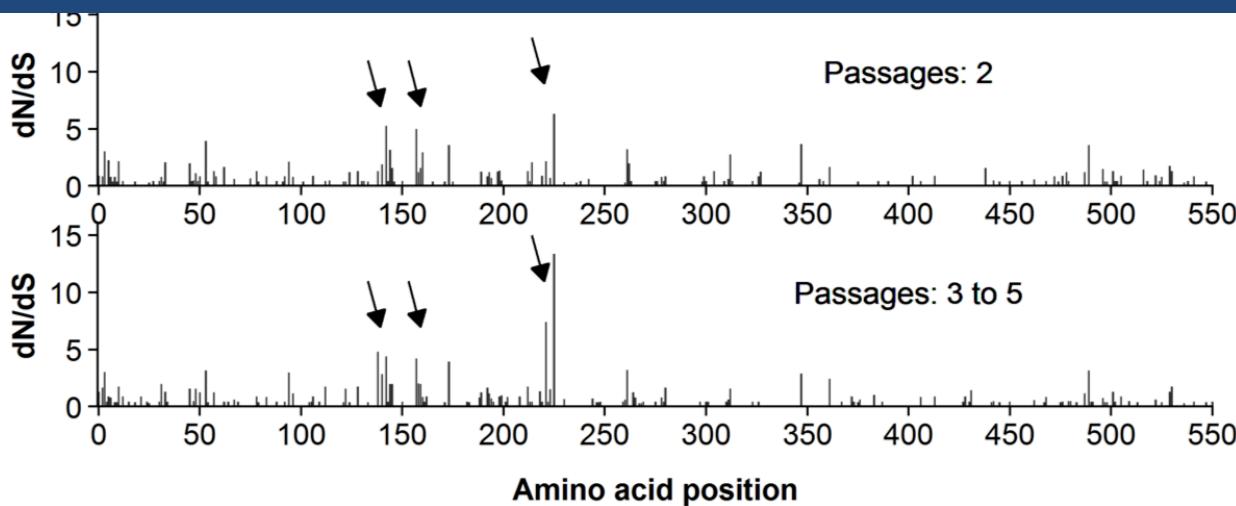
unpassaged virus



Signal increases with the number of passages

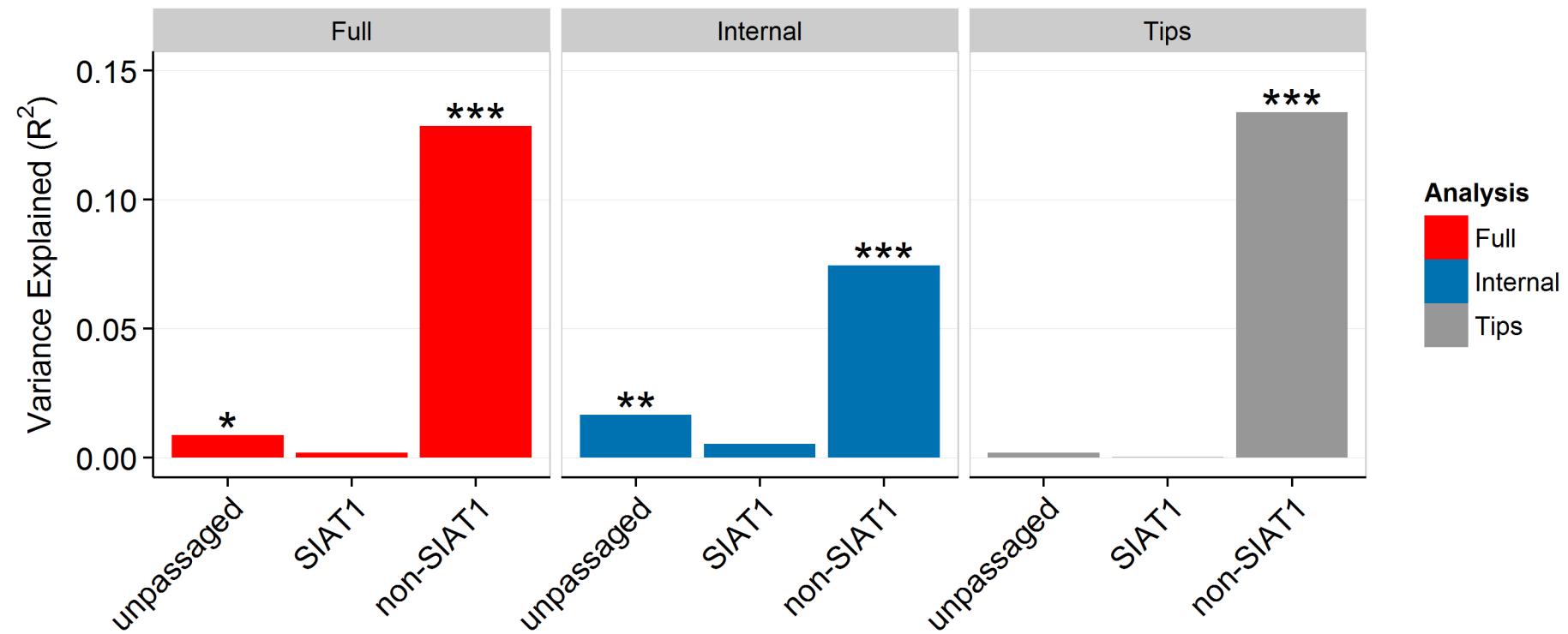


Signal increases with the number of passages



Virus passaged in MDCK-SIAT₁ cells looks like unpassaged virus

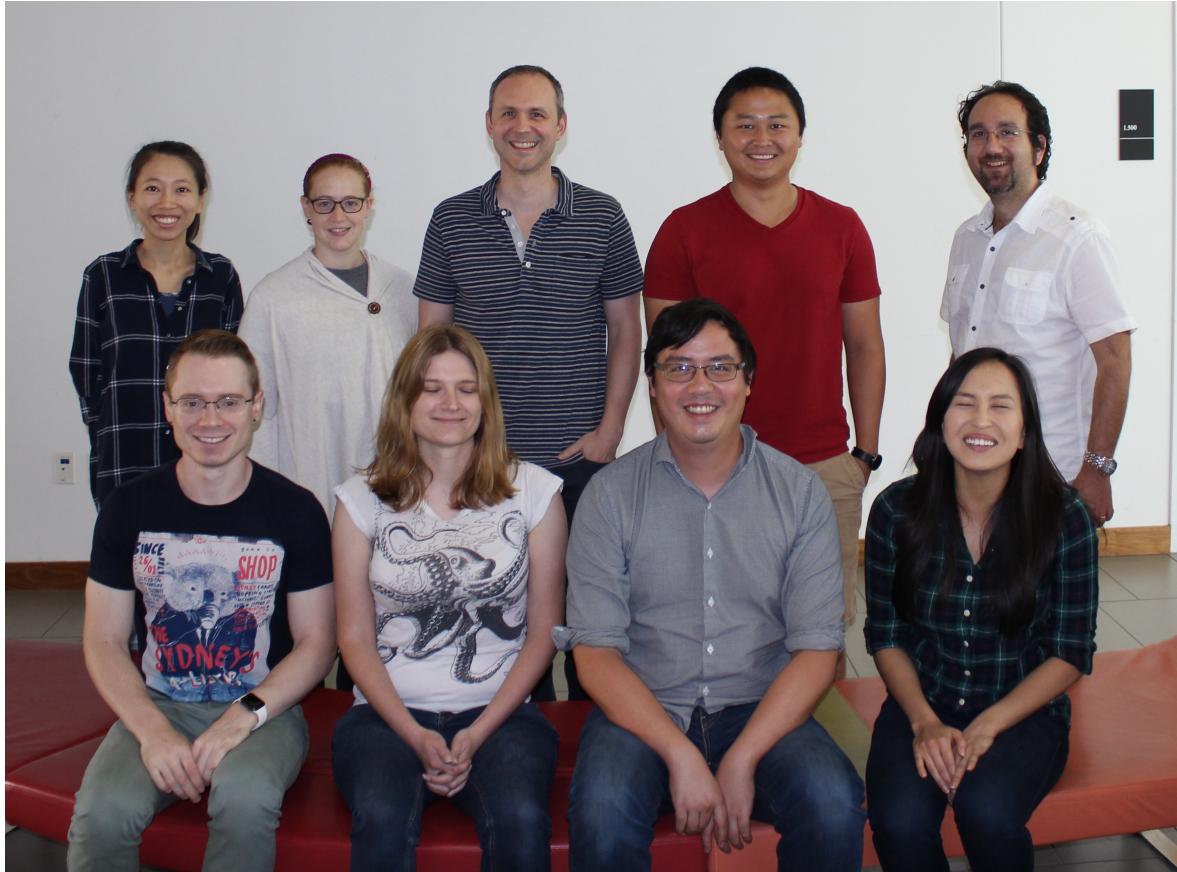
1046 sequences per group



Take-home messages

- Influenza virus is regularly adapted to cell culture, with the purpose of measuring its antigenic profile
- As a result, analyzed viral strains are often not representative of the circulating virus

Acknowledgments



Collaborators

- Julian Echave
- Ed Marcotte
- Claire McWhite
(Marcotte lab)
- Austin Meyer

Funding

- NIH/NIGMS
- NIH/NIAID
- Army Research Office
- NSF BEACON Center